ORIGINAL RESEARCH **Open access**

Robust Object Tracking for UAVs via Event-Guided RGB Enhancement in High Dynamic Range Scenes

Hiroshi Nakamura[1*], Yuta Kato[1], Rina Okabe[2], Kenji Mori[3]

Abstract

Unmanned aerial vehicles have been deployed across numerous applications, including atmospheric observation, firefighting and rescue, and agricultural spraying, where precise environmental awareness is essential for both decision-making and control. Among onboard sensors, the RGB camera is vital for acquiring detailed visual data needed for navigation. Yet, it faces a major difficulty in high-dynamic-range conditions, which are common in real-world operations. In these situations, captured frames simultaneously exhibit under-illumination and over-illumination, degrading downstream vision tasks. To address this issue, we use object tracking as a representative case and exploit event cameras' superior response across a broad intensity range to introduce an event-assisted object-tracking method that is robust to extreme intensity changes. Concretely, we propose extracting feature correspondences from dense event streams and, building on this, (i) develop a U-Net-based image enhancement technique to equalize RGB intensity by utilizing temporally adjacent frames, and subsequently (ii) build a dual-input tracking model to follow moving objects using both intensity-equalized RGB video and event sequences. The proposed framework is thoroughly evaluated through both simulated and real-world experiments.

Keywords Drones, Harsh illumination, Image enhancement, Event-Assisted object tracking, Multi-Sensor fusion

*Correspondence:
Hiroshi Nakamura
nakamura.hiroshi@agr.nagoya-u.ac.jp
[1] Department of AI for Mushroom and Wasabi Cultivation Environment Control, Graduate School of Agricultural Sciences, Nagoya University, Nagoya, Japan
[2] Department of Spectral Analysis for Nitrogen Deficiency Detection, Faculty of Agriculture, Keio University, Tokyo, Japan
[3] Department of Smart Harvesting Automation, Faculty of Agriculture, Tohoku University, Sendai, Japan

Introduction

Owing to their lightweight build, agility, and affordability [1-3], drones have become commonplace in numerous remote activities such as patrol [4, 5], reconnaissance [6], and cargo distribution [7]. During these missions, drones need to interpret their surroundings to inform subsequent actions and decisions correctly. Standard sensing payloads on UAVs include visible-band optical cameras [8], LiDAR [9], near- and mid-infrared imagers [10], and so forth. Every sensing modality possesses distinct merits and limitations, rendering multi-modal perception the prevailing paradigm in this area. Amid the suite of sensors, the visible-band camera constitutes a non-negotiable perceptual element thanks to its fine spatial detail, information density, and modest build cost.

As a fundamental UAV competency, object tracking [11-14] has drawn extensive research interest. Broadly, tracking algorithms ingest either the RGB stream or its fusion with auxiliary modalities. Purely RGB-driven trackers [15-18] perform well in conventional frame-based settings but break down under harsh lighting. Several investigators have sought to assimilate cues from event-driven cameras, which thrive in both dim and high-contrast environments. To unify RGB

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

frames and event flows, Mitrokhin *et al.* [19] offered a time-image descriptor to condense temporal characteristics of the event stream, and Chen *et al.* [20] advanced event encoding by devising a synchronized Time-Surface with Linear Time Decay descriptor. These strategies yield encouraging tracking outcomes while preserving strong temporal continuity.

Nonetheless, the foregoing techniques struggle to transfer to day-and-night UAVs because RGB sensors have a limited perceptual range in complex lighting conditions. Seeing that overexposure and underexposure alike gravely corrupt image fidelity and obstruct accurate tracking, trustworthy aerial sensing of harshly illuminated environments stays profoundly difficult [21, 22]. Consider the footage in **Figure 1**: while filming a vehicle traversing a tunnel, every frame exhibits a broad luminance spread and sudden inter-frame variations; the vehicle essentially vanishes in certain frames for both trackers and the human eye due to severe underexposure. Fortunately, drones inherently experience continually shifting illumination when airborne, producing recordings of non-uniform quality over a target region. Given the predominantly stable feature points across consecutive video frames [23, 24], we are prompted to rectify poor-quality frames using their high-quality correspondents, achieving consistently high-quality video and resilient downstream functions. One essential hurdle is identifying corresponding features across successive frames that are subject to sudden luminance jumps.

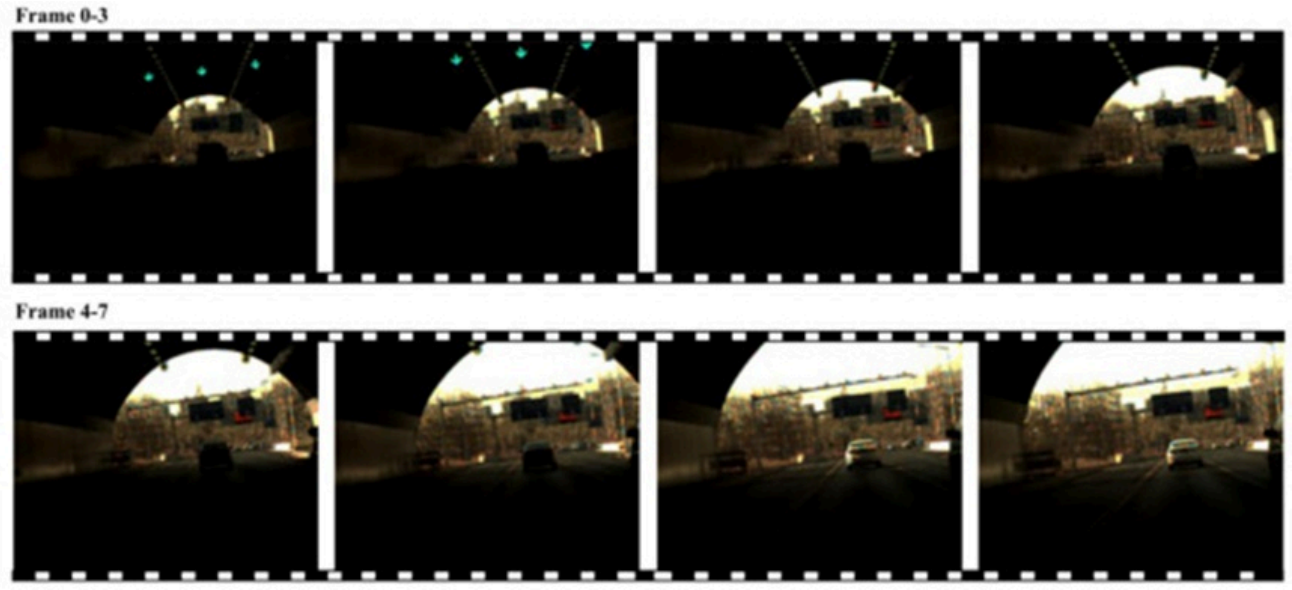


Figure 1. A representative high-dynamic-range RGB sequence of a car passing through a tunnel, where the vehicle is barely discernible during the last 5∼6 frames owing to underexposure.

To improve image quality under harsh lighting, considerable research has been conducted in recent years. A widely adopted approach is to recover HDR outputs by fusing a stack of differentially exposed LDR captures [25]. In moving settings, alignment is essential to resolve discrepancies across frames with dissimilar exposures. Kang *et al.* [26] initially co-registered surrounding frames against a target and blended the aligned outputs to synthesize an HDR frame. Later studies [27, 28] augmented this by inserting a motion estimation stage and a refinement pass. Charting another course, Kalantari *et al.* [29] formulated a patch-wise optimization approach that synthesizes missing exposures inside each frame before reconstructing the final HDR result. Gryaditskaya *et al.* [30] further improved this pipeline by incorporating an adaptive metering mechanism that adjusts exposure settings, thereby reducing motion-induced artifacts. Instead of capturing multi-exposure stacks, some methods use deep neural networks to reconstruct HDR imagery from a single input. However, hinging on a static reference exposure renders the inversion strongly ill-posed and unable to preserve robust frame-to-frame coherence.

Furthermore, many established HDR video reconstruction pipelines concentrate on engineering specialized hardware, e.g., scanline exposure/ISO [31-33], per-pixel exposure [34], and modulo cameras [35], yet these emerging cameras

© 2023 The Author(s). | Published by Agro Intelligence Research Press

remain investigational and not market-ready in the foreseeable future. Another line of recent work operates within the deep-optics paradigm, jointly optimizing the optical front-end and a CNN-based decoder for HDR imaging challenges. The aforementioned approaches typically embed presuppositions about scene illumination that may collapse in unconstrained environments. Besides, a majority of these algorithms require ground-truth HDR exemplars for supervised training, and exhibit diminished effectiveness in scenes that deviate from the training distribution. Hence, these methods are conceptually instructive but difficult to field directly on operational UAV platforms roaming open settings.

The event camera, also known as a neuromorphic retina, is a nascent sensing modality that asynchronously reports intensity changes that cross a threshold [36, 37]. In recent times, event signals have fueled multiple high-speed applications capitalizing on their acute sensitivity and near-instantaneous response, spanning high-speed tracking [38-41], frame interpolation [42, 43], optical flow computation [44-46], and motion detection [47]. Unlike frame-based optical sensors, event cameras output “events” signaling that a sufficiently large luminance shift has transpired at specific pixels and timestamps, along with the polarity of the change. Observing that an event camera records motion across a wide luminance range and remains insensitive to abrupt intensity transitions, we advocate leveraging event signals to explicitly align RGB frames, thereby mitigating the quality degradation detrimental to successive object tracking. Stated another way, drawing on the consistent portrayal across event streams, we elevate low-grade frames under the guidance of their high-grade counterparts, thereby maintaining continuous high-grade scene awareness. In practice, we correspond key points at different timestamps [48] and exploit this correspondence to counterbalance luminance swings across consecutive RGB frames. Afterwards, we establish a fusion backbone that blends the rectified RGB frames with event data for reliable object tracking.

The contributions of this paper are as follows:

· We propose an event-guided, resilient object-tracking framework tailored for high-dynamic-range environments, which effectively fuses inputs from an event camera and an RGB camera to mitigate the destructive effects of hostile lighting on tracking performance. To the best of our knowledge, this marks the first investigation into object tracking under severe illumination conditions using dual-modal cameras.

· We construct an end-to-end deep neural architecture that enhances high-dynamic-range RGB frames and performs object tracking sequentially, trained in an unsupervised regime. Quantitative assessments indicate that the proposed solution raises tracking precision by up to 39.3%.

· We develop a technique for aligning feature points that manifest at distinct time instances within the dense event stream, thereby steering luminance rectification in high-dynamic-range RGB frames. The devised feature alignment can register key points across high-dynamic-range frames spanning a 1-second window.

· The approach demonstrates exceptional performance across a spectrum of harshly lit conditions, corroborating its effectiveness and substantially extending the operational reach of drones.

In Section 2, we first outline the framework and algorithm design of the proposed event-assisted object tracking, encompassing event-driven cross-frame alignment, RGB image enhancement, and dual-modal object tracking. Section 3 details the experimental configuration, including datasets and training methodology. The qualitative and quantitative findings are examined in Section 4. Additionally, we showcase the results on real-world recordings together with the ablation study. Finally, Section 5 concludes the paper, discusses the limitations of the proposed method, and spotlights future research directions for efficient cooperative sensing around UAVs.

Framework and Algorithm Design

This section outlines the specifics of the proposed event-assisted robust object-tracking scheme for harsh illumination conditions. We begin with a concise overview of the overall architecture, followed by an in-depth description of three

© 2023 The Author(s). | Published by Agro Intelligence Research Press

pivotal modules: the extraction of cross-frame feature correspondences, the restoration of high-dynamic-range frames, and the subsequent dual-modal object tracking.

The core principle behind the proposed scheme is to leverage event cameras' reliable motion perception capabilities to prevent degradation of RGB frame quality and, thereafter, merge event cues with the restored RGB footage to reinforce tracking performance compromised by overexposure and underexposure. The full architecture of the proposed event-assisted object-tracking scheme is shown in **Figure 2**; it comprises three principal modules.

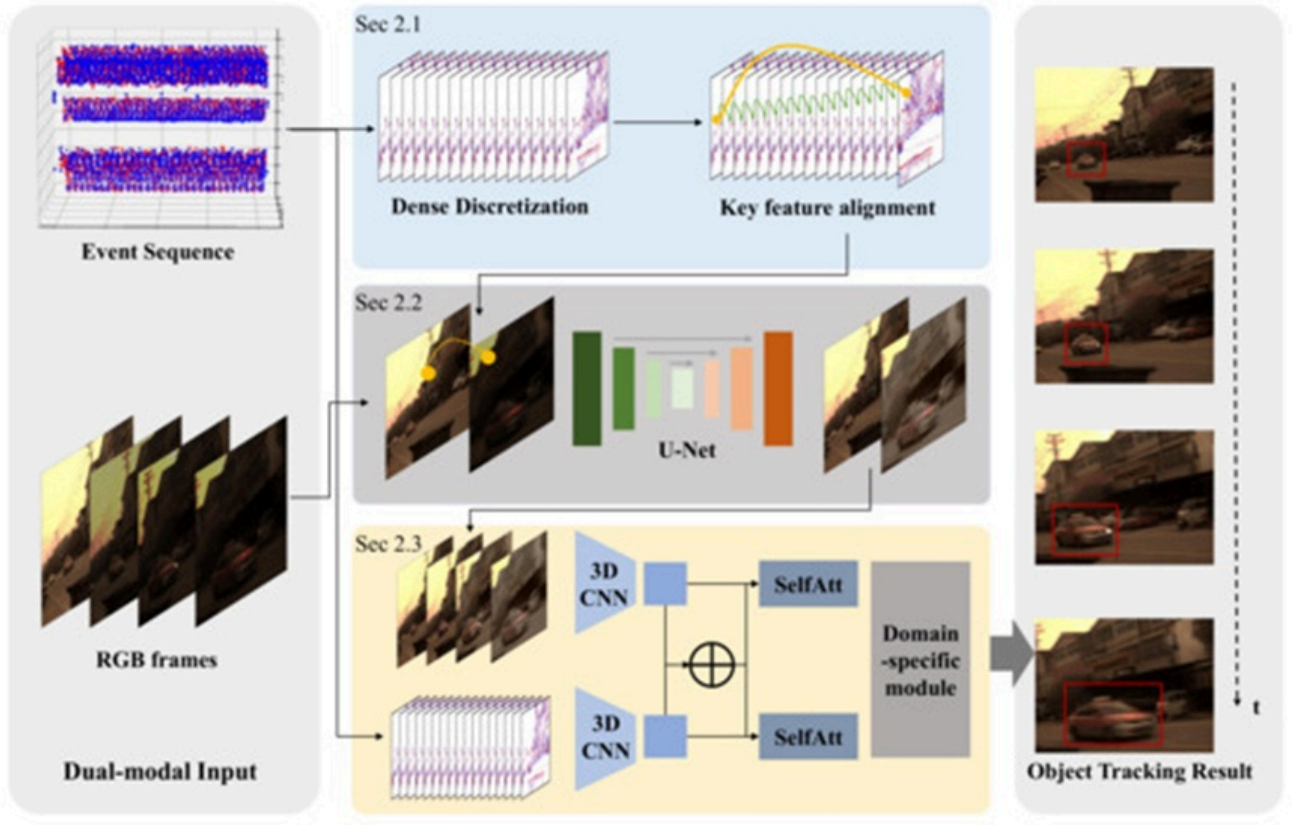


Figure 2. The architecture and processing pipeline of the event-assisted robust tracking algorithm under harsh illumination. The entire pipeline runs fully automatically and comprises three critical phases: the first relies on traditional optimization, and the latter two use deep neural networks.

(i) Reconstructing the displacement paths of salient feature points from the dense event stream. We segment the event stream into batches that fall within overlapping, short-duration time bins, and the salient points obtained from Harris corner detection within each batch can then form displacement paths. We then stitch these short, locally defined paths to derive motion spanning a broader time interval across the RGB frames, even in the presence of severe illumination changes.

(ii) Restoring the high-dynamic-range RGB frame through cross-frame alignment and information sharing. Using correspondences among feature points across frames, we construct a deep neural network to repair overexposed or underexposed regions by drawing on adjacent frames with superior visibility to guide low-visibility target frames. In practice, we adopt a U-Net-based neural network for this image restoration task.

(iii) Tracking target objects by fusing cues originating from both RGB and event modalities. We devise a tracker that accepts two-stream inputs, merging information from the restored RGB frames and event sequences to determine object

© 2023 The Author(s). | Published by Agro Intelligence Research Press

motion paths. Specifically, we deploy 3D CNNs to extract features, blend the features from the two processing branches through a self-attention module, and finally use an MLP to predict the resulting object motion.

Event-based cross-frame alignment

Event-driven keypoint extraction and matching are performed at this stage to leverage illumination-invariant event signals for cross-frame alignment of the degraded RGB footage, enabling frame repair using spatially corresponding patches of adequate quality in neighboring frames. We localize keypoints of moving objects within the event stream, as depicted in **Figure 3**.

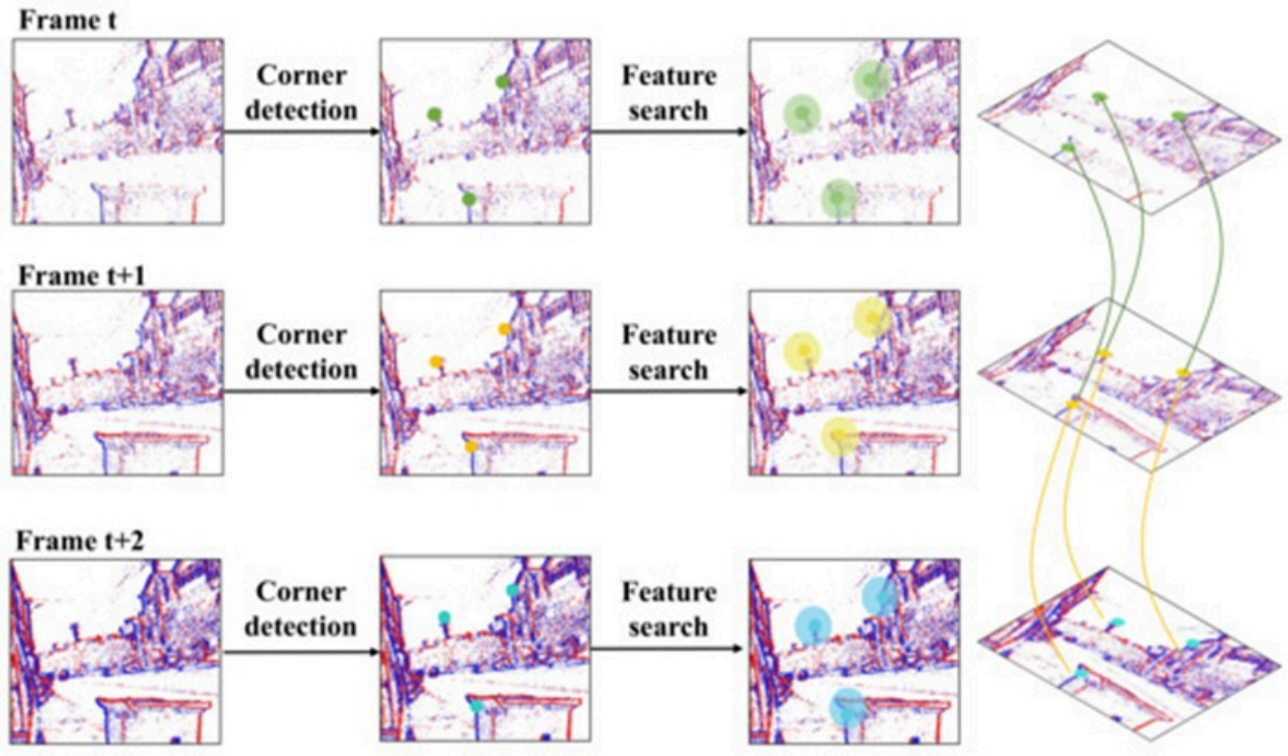


Figure 3. Illustration of event-based keypoint alignment. Keypoints are detected through the Harris operator, and locally matching counterparts are sought (circular search zones are marked with distinct colors) to assemble motion paths, as shown in the right column.

Assuming a time span T, we consider N RGB frames and S event signals. We label the collected RGB frames as $\{I0,1,\cdots,IN\}$, with their respective capture instants denoted by $\{T0,T1,\cdots,TN\}$. An individual event signal s is expressed as a quadruple $(xs,,,)$, where xs,ys specify the pixel coordinates of s; ts gives the firing time; and Ps encodes the polarity of the brightness change. Initially, we divide the S event signals into $K\times N$ slices along the temporal axis and collapse each slice into a 2D projection, which we term an event frame, yielding $K\times N$ such frames. The Harris corner detector is applied to these event frames to locate distinct keypoints. Next, we establish correspondences between keypoints across different time instants. Presuming the shape of the moving object stays approximately rigid within a very short interval—meaning keypoint configurations in successive frames are close—we define a compact circular search neighborhood of radius r around each keypoint. That is, a keypoint in the e-th frame is paired with features lying inside the search neighborhood of the $(e+1)$-th frame.

For the n-th RGB frame, we initially bring the event frames ranging from $n\times S$ to $(n+1)\times S$ into alignment. The displacement measured between the features across several event frames allows one to trace out a motion path of the event keypoints,

© 2023 The Author(s). | Published by Agro Intelligence Research Press

which reflects the displacement of the matching keypoints within the RGB frame. By extension, we can ultimately deduce the location of the matching keypoint when moving from the n-th to the (n+1)-th RGB frame.

RGB image enhancement

After aligning feature points across different RGB frames, we repair underexposed and overexposed regions by leveraging their high-visibility correspondences to recalibrate brightness and restore missing details. To keep the description concise, we refer to the low-visibility frames as targets and the high-visibility frames as references. Achieving this restoration requires addressing two central issues: (i) how to determine which target frames demand enhancement; (ii) how to architect a learning model capable of lifting visibility to align with the reference frame while safeguarding the native structure of the target frame.

We begin by gauging frame visibility to identify heavily degraded frames. Intuitively, because harsh illumination produces local over- or underexposure, which typically manifests as a scarcity of texture, we adopt information richness as a proxy for degradation severity. In our implementation, visibility (Vi) of an input RGB image Ri is quantified as the discrepancy measured against its low-pass filtered variant ($\hat{R}\ i$), namely,

$$V_t = Var\left(R_t - \hat{R}_t\right), \quad (1)$$

Where (·) signifies variance computation.

In general, we partition frames into groups and perform compensation within each group. We repeatedly pick the target frame with the lowest visibility score and the reference frame with the highest visibility, then carry out enhancement. The loop terminates when the iteration count exceeds a preset limit P or when the visibility gap between target and reference falls below η. In our tests, we set P=10 and η=0.1.

For the enhancement step, we engineered a U-Net-style network that predicts the restored frame given the target and reference frames, as portrayed in **Figure 4** [49]. The network consists of a three-stage encoder for feature extraction and a three-stage decoder. Skip connections are introduced to help preserve spatial detail. The network is learned without supervision. We formulate the loss around the aligned keypoints. Recognizing that the restored frame ought to resemble the reference near keypoints while remaining faithful to the original elsewhere, we assemble a composite loss function. To promote the former similarity, we minimize the MSE distance; for the latter, we use the LPIPS metric. Denoting the reference by Iref, the original target by Iobj, and the network output by Iout, the loss is expressed as:

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

Where k indexes the keypoint positions and $\neg k$ covers the remaining pixels, the hyper-parameter α is held at 0.05 throughout training.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

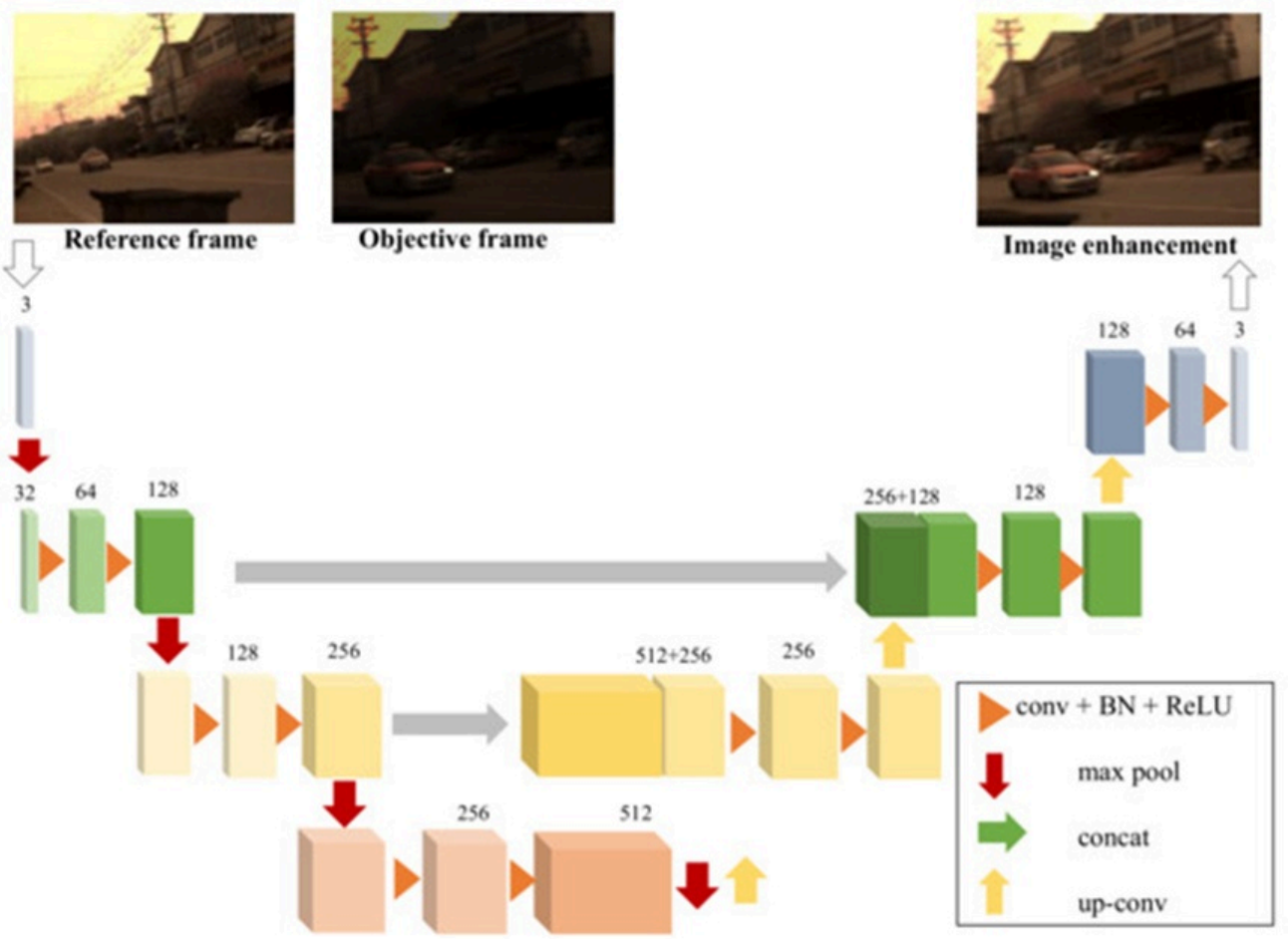


Figure 4. Structure of the RGB image enhancement module. The captured RGB frame is passed to a U-Net comprising a three-stage encoder for feature extraction and a three-stage decoder for image enhancement. Skip connections bridge the corresponding encoder and decoder stages to help retain spatial information. The diagram shows convolutional, pooling, upsampling, and downsampling layers, with key operations labeled as follows: conv for convolution; BN for batch normalization; ReLU for the ReLU activation; max pool for max pooling; concat for concatenation; and Up-conv for transposed convolution.

Dual-mode object tracking

To leverage motion cues in both the event stream and the restored RGB frames, we assemble a dual-mode tracking component for robust object tracking. This component adopts RT-MDNet [50] as its backbone. It contains a shared feature-mapping subnetwork that builds a unified representation that separates the target from the background, along with a domain-specific subnetwork that focuses on extracting information invariant across domains. Unlike the original RT-MDNet [50], the dual-mode design proposed here emphasizes integrating features from two modalities. It embeds two self-attention blocks to amplify the combined representation derived from the two input branches.

The network layout is shown in **Figure 5**. First, we set up two independent 3D CNN streams to extract features from the respective inputs and deliver feature vectors of identical dimensions. These two vectors are then concatenated and processed through a convolution operation to yield a fused feature representation. After that, we build the self-attention network to surface the information embedded within each standalone feature input [51]. A fully connected MLP with two layers is employed to produce the shared feature. Following the approach of RT-MDNet [50], we construct a domain-specific layer that ultimately generates the tracking outputs.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

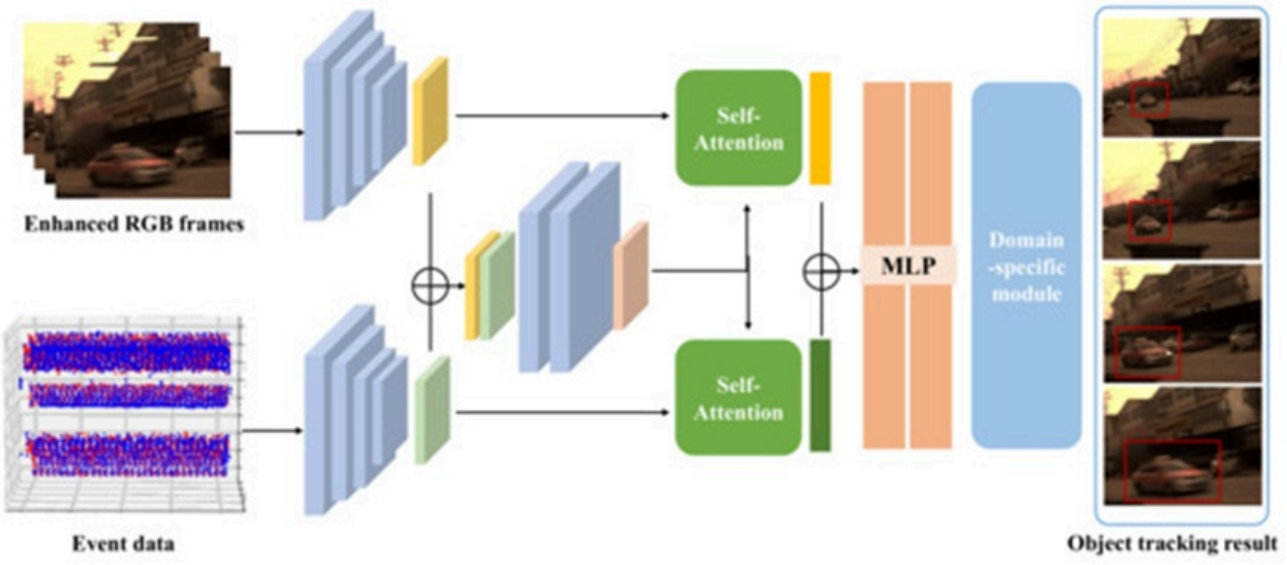


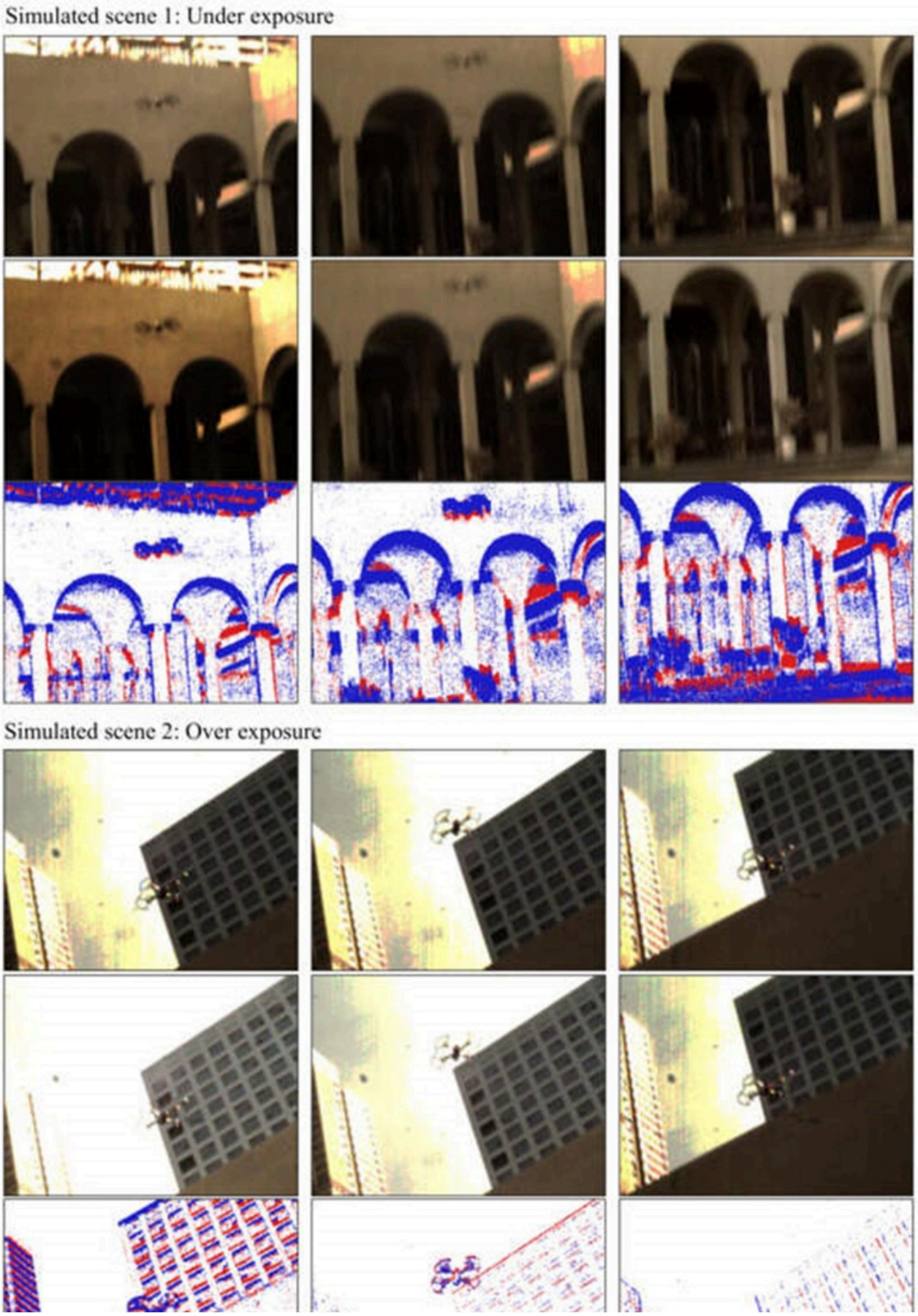
Figure 5. Layout of the object-tracking module. The RGB frames and the event sequence are each fed into separate 3D CNN blocks for feature extraction; the resulting features are concatenated and forwarded to another CNN block for fusion. Then, the individual and fused features are separately passed into the self-attention network. Finally, two MLPs are utilized to produce the object detection and tracking outputs.

During training, a cross-entropy loss is computed for each detection bounding box to maximize the separation between the foreground target and the background, a principle also applied across multiple domains. In the fine-tuning phase that follows, separate strategies are adopted for the initial frame and the remaining frames of a sequence. For the very first frame, we sample multiple bounding boxes from a Gaussian distribution to enable domain-specific adaptation. In contrast, for all subsequent frames, we generate random samples conditioned on the outcome of the preceding frame and determine the optimal bounding box through regression.

Experimental Settings

· Datasets: We assess the proposed method using both synthetic and real-world data. VisEvent [52] serves as the source for the synthetic data, where harsh illumination effects are introduced by manipulating the brightness and contrast of the RGB frames. In detail, the manipulations are implemented as follows: luminance is altered to vary linearly, quadratically, or exponentially over the sequence, whereas image contrast is adjusted linearly with varying rates of change. We begin by randomly selecting one-third of the data for luminance manipulation and subsequently apply contrast manipulation to one-third of the randomly chosen videos. Two examples extracted from the synthetic dataset are depicted in **Figure 6**. The first scene reproduces the brightness shifts typical of underexposed conditions, and the second replicates overexposure, both achieved through modifications to image brightness and contrast. As shown, starting from original clips with nearly uniform illumination, we can generate footage under complex lighting conditions. In the resulting high-dynamic-range RGB frames, textured areas within certain regions become indiscernible in some frames due to underexposure or overexposure. Meanwhile, contour information across the entire field of view is reasonably well preserved.

© 2023 The Author(s). | Published by Agro Intelligence Research Press



© 2023 The Author(s). | Published by Agro Intelligence Research Press



in luminance to replicate the overexposure present in its first frame.

Regarding the real-world recordings, we collected a set of representative nighttime traffic scenarios using a pair of spatially registered cameras (one conventional RGB and the other event-based). These scenarios feature complex illumination (e.g., traffic lights, glowing signs) and substantial intensity swings. The two representative examples in **Figure 7** reveal that such settings are characterized by wide luminance fluctuations and that the traffic participants become nearly unidentifiable in certain frames, whether due to underexposure or overexposure. This difficult dataset offers a direct testbed for evaluating the effectiveness of the proposed object-tracking algorithm in practical environments, as shown in **Figure 7**.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

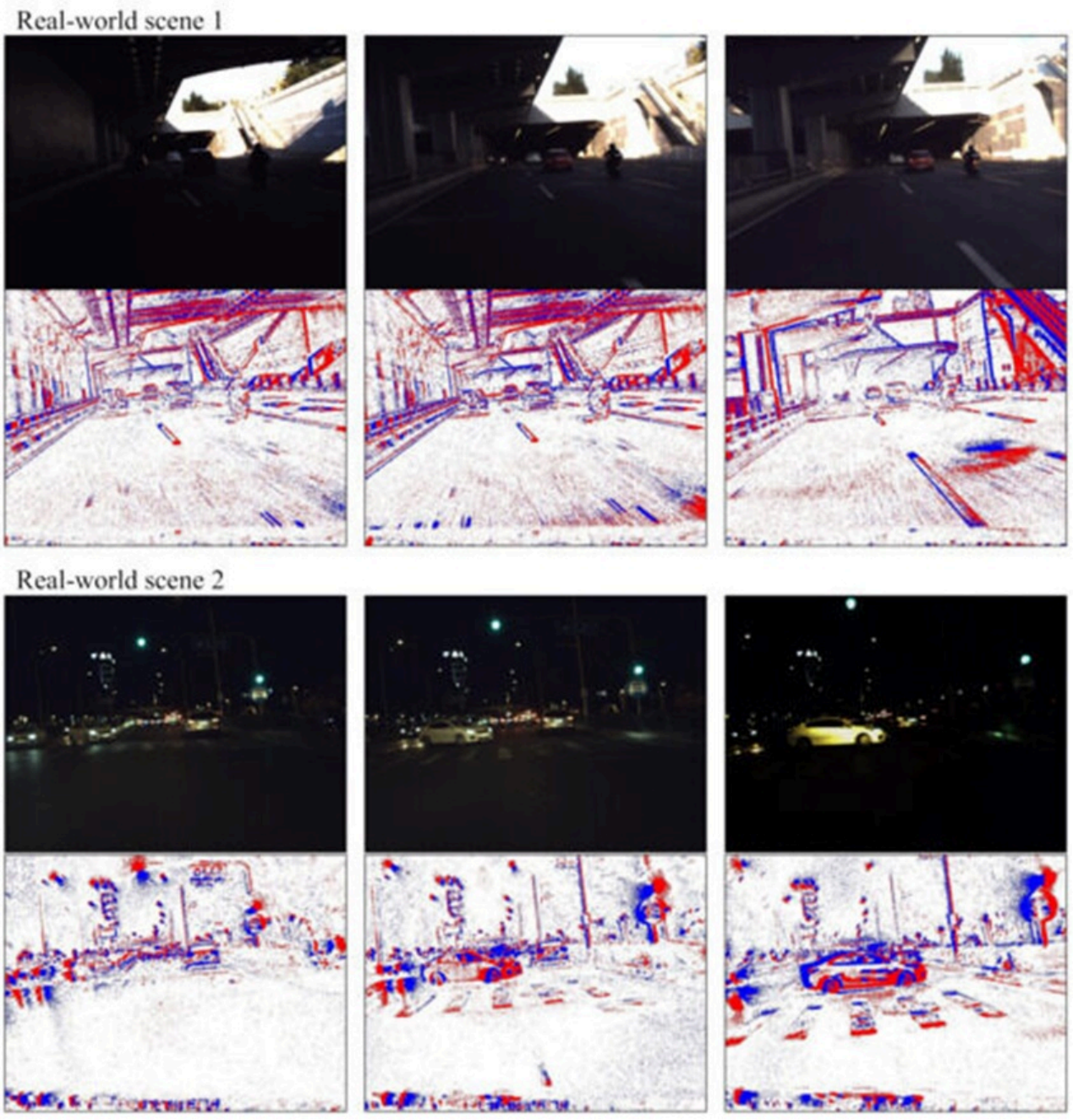


Figure 7. Two representative instances from the real-world dataset were recorded in harshly lit traffic settings: one beneath a bridge during the day and another at a nighttime intersection. In each case, the RGB and event cameras have been pre-calibrated to ensure pixel-wise correspondence.

· Baseline algorithms: Three state-of-the-art tracking algorithms are selected as baselines against the proposed solution, namely RT-MDNet [50], Siamrpn++ [53], and VisEvent [52]. Both RT-MDNet [50] and Siamrpn++ [53] are RGB-input trackers that achieve strong performance under standard lighting conditions. Given that very few dedicated algorithms exist for harsh illumination conditions, we selected the aforementioned two robust and popular trackers for comparison. VisEvent [52] implements a two-modality neural network that merges RGB and event information; we benchmark the proposed solution against VisEvent [52] to assess the contribution of the image enhancement module under unfavorable illumination conditions. This reference method accepts inputs comparable to ours and delivers leading-

© 2023 The Author(s). | Published by Agro Intelligence Research Press

edge performance, making it a suitable candidate for verifying the proposed image enhancement module under severe lighting conditions.

· Training: The training procedure runs on an NVIDIA 3090 and takes approximately 4.7 hours. Both the input image dimensions and the spatial resolution of the event sequence are set to 640×480 pixels, and seven successive RGB frames (∼350 ms) are used for intensity equalization. Optimization is performed with the Adam algorithm, using a learning rate of 5×10^{-4}, momentum of 0.9, and a weight decay of 5×10^{-4}.

Results and Discussion

Within this section, we design a series of experiments to confirm the effectiveness of the proposed method on two distinct tasks in high-dynamic-range environments: image enhancement and object tracking. We begin by reporting the visual and quantitative performance relative to the chosen baseline algorithms. Subsequently, we present qualitative results on real-world data to highlight further the visual contrast between the proposed solution and the baselines. Finally, we conduct ablation experiments to measure the contribution of the algorithm's core module.

Results based on simulated data

In the present subsection, we evaluate our approach with respect to both image enhancement quality and object-tracking accuracy, using the synthetic data. Both qualitative and quantitative experimental findings are supplied to permit a thorough analysis of the proposed solution's effectiveness. For the qualitative portion, we first display the image enhancement results and then compare the object-tracking performance with that of the baseline algorithms. For the quantitative portion, we compare the precision plot (PP) and success plot (SP) metrics when assessing the tracking capability.

Qualitative results

Figure 8 presents qualitative outcomes for a sample video drawn from the synthetic dataset. The top row shows the raw RGB sequence, characterized by large intensity disparities within individual frames and across the sequence. In this scene, a person sprints from a brightly lit region toward a destination shrouded in deep shadow. Owing to the extremely dim conditions, discerning the person's silhouette in the ultimate frame proves difficult. We restore the RGB frames using temporal correspondences derived from the event signals, as shown in the middle row. The restored variant achieves a substantially more uniform intensity distribution and brings out the human outline even under very weak illumination.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

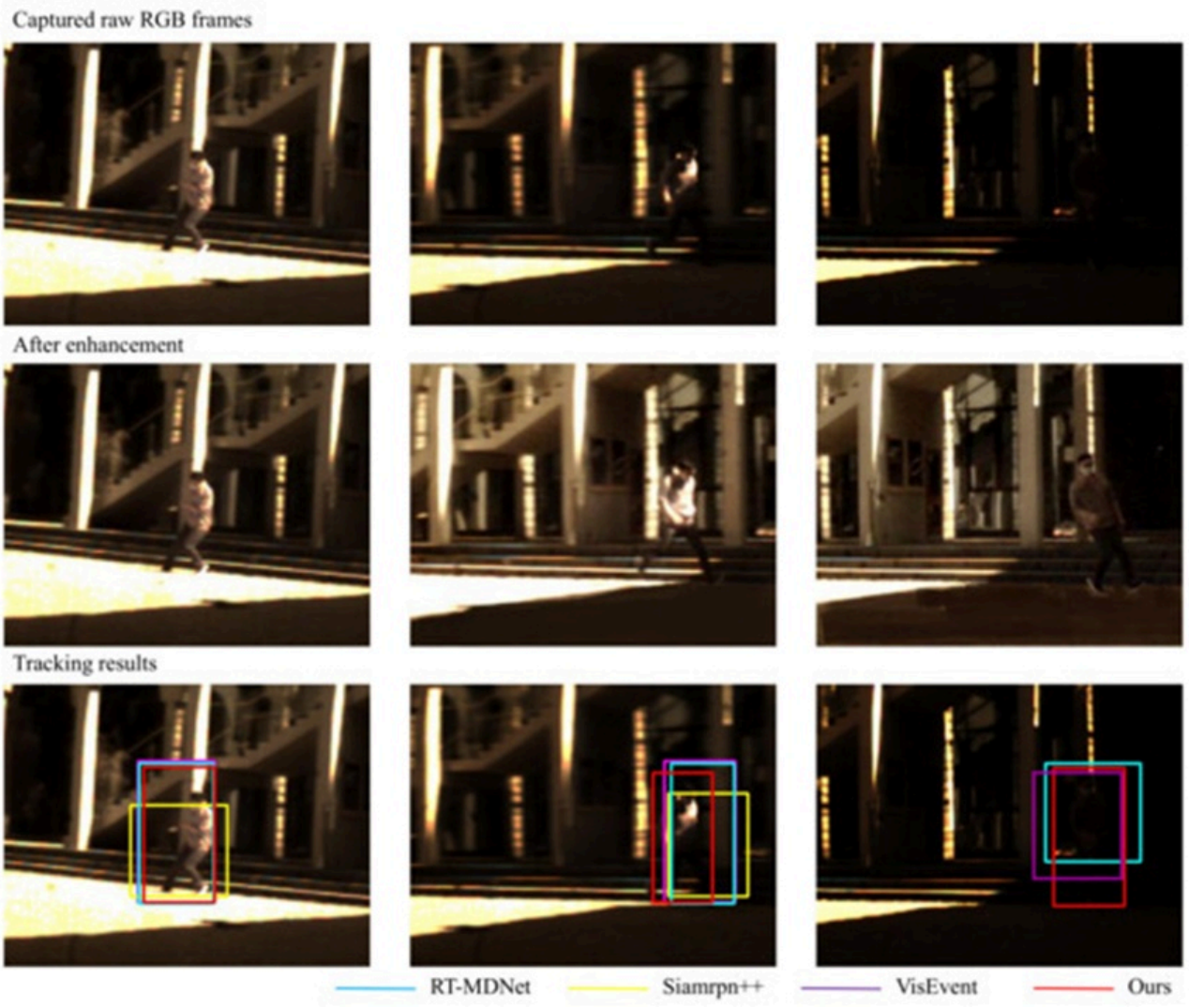


Figure 8. Illustration of RGB enhancement performance within object tracking for a typical example scene from the synthetic dataset. (Top) The originally captured RGB video frames. (Middle) The enhanced images produced by the proposed method are matched. (Bottom) The tracking results obtained by different object-tracking algorithms.

The bottom row additionally illustrates the object-tracking outputs. The bounding boxes generated by our approach and by the three competing methods are overlaid and color-coded. In sufficiently lit conditions, every algorithm can track the object with high precision. The performance of RT-MDNet, VisEvent, and the proposed algorithm is roughly on par, whereas the bounding box produced by the Siamrpn++ tracker exhibits some displacement. As the light level drops, the proposed algorithm continues to locate the person, while the bounding boxes from RT-MDNet and VisEvent begin to drift. When light becomes extremely scarce, only the proposed method, RT-MDNet, and VisEvent can still stay on the object, thanks to the high sensitivity of event signals and their tolerance to abrupt brightness changes. In contrast, the RGB input used by RT-MDNet and VisEvent is not enhanced, which lowers their final tracking accuracy, whereas our approach consistently maintains reliable tracking.

Quantitative results

We employ the standard evaluation measures PP and SP to gauge object-tracking precision. Concretely, PP quantifies the percentage of frames in which the displacement between the estimated object center and its ground-truth position stays within a preset threshold. SP measures the percentage of frames in which the IoU between the predicted bounding

© 2023 The Author(s). | Published by Agro Intelligence Research Press

box and the ground-truth annotation exceeds a preset threshold. **Table 1** reports the PP and SP figures attained by our method, together with three top-performing object-tracking algorithms. Due to the absence of ground-truth labels for real-world footage, quantitative assessment is confined to synthetic data.

Table 1. Quantitative comparison across different object-tracking algorithms on the synthetic dataset, expressed through PP and SP.

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

As shown in **Table 1**, the proposed algorithm achieves the highest tracking precision. Under severe illumination, our method maintains a steady lock on the target, whereas Siamrpn++ and RT-MDNet exhibit substantially degraded performance under the same conditions. Furthermore, although VisEvent incorporates event signals into its input, it disregards the corrupting influence of poor-quality RGB frames, thereby achieving lower tracking fidelity. From the ranking, we can distill two key insights: first, event signals demonstrably help mitigate performance erosion in high-dynamic-range situations; second, restoring the impaired RGB frames provides an additional lift to object-tracking accuracy.

Results based on real-world data

To probe how our approach behaves in authentic high-dynamic-range settings, we run the algorithm on several video clips collected under demanding lighting, with one typical case displayed in **Figure 9**. This footage was shot at the mouth of a tunnel, and the top-row frames show a car passing through it. Once the car enters the tunnel, obtaining clear visual images becomes difficult due to insufficient lighting, and the car becomes virtually unnoticeable in the last frame. The middle row presents the restored outputs, indicating that the perceptual quality of the RGB frames is substantially upgraded relative to the raw capture.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

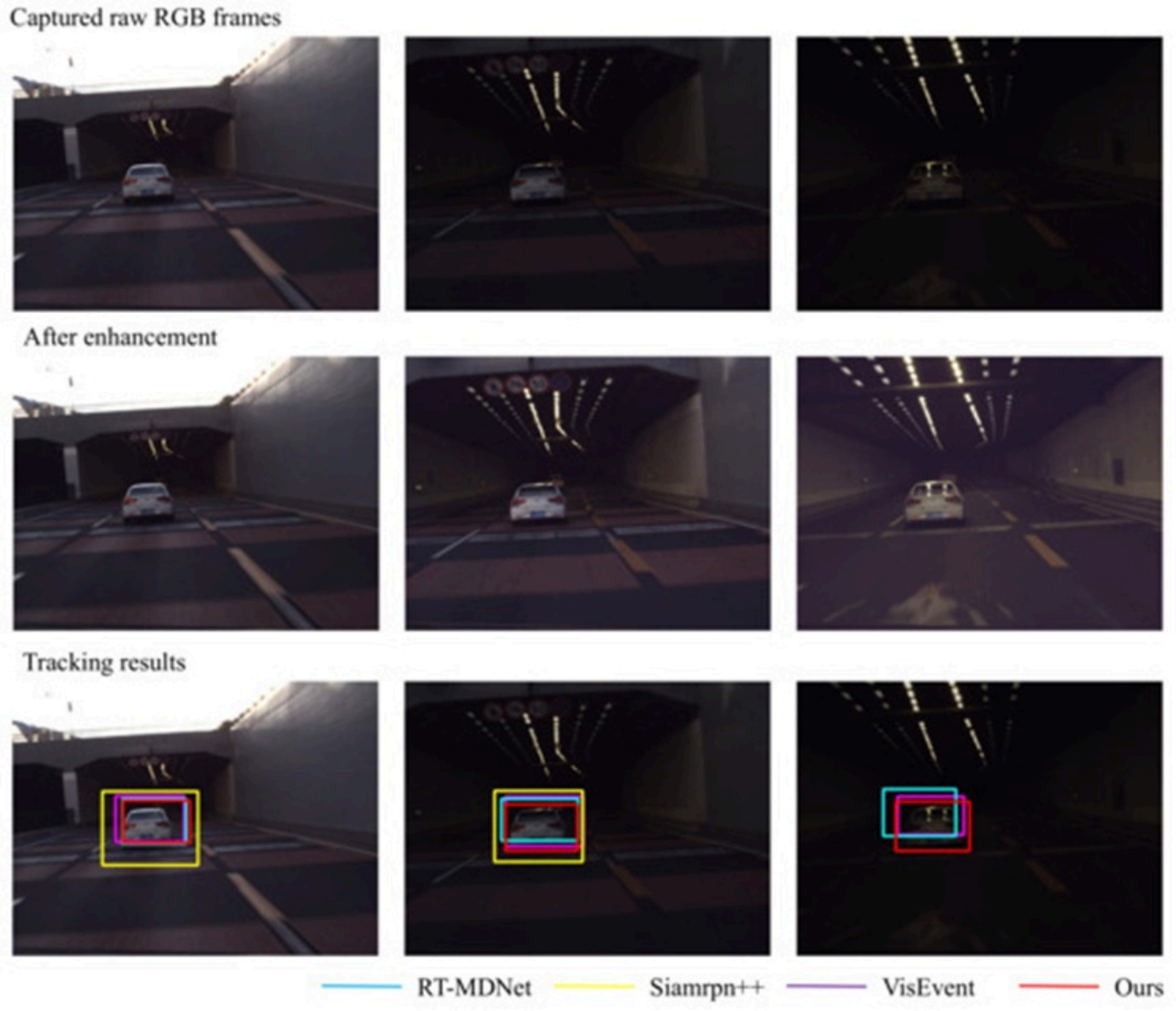


Figure 9. Demonstration of our image enhancement on the tracking output and side-by-side performance evaluation against existing object-tracking algorithms on a real-world high-dynamic-range scene—a white car driving through a tunnel: (Top) The raw captured RGB frames, (Middle) Our restored RGB images, and (Bottom) The tracking outcomes produced by different algorithms.

The tracking outputs are shown in the bottom row of **Figure 9**. All four algorithms successfully follow the car when the lighting is ample. Once illumination weakens, the two trackers that rely exclusively on RGB input falter: Siamrpn++ loses the car altogether, and RT-MDNet generates a bounding box that drifts considerably; VisEvent, on the other hand, proves comparatively more stable but still produces imprecise bounding boxes. By contrast, our method sustains trustworthy tracking over the full sequence. These experimental observations reinforce two points: (i) lighting conditions strongly influence object-tracking quality, and (ii) event signals can aid object tracking when illumination is unfavorable.

Ablation studies

The ablation study aims to quantify the contribution of the event-driven temporal alignment step to both RGB image restoration and object tracking. Within the proposed pipeline, we rely on Harris corner detection to extract salient keypoints from the dense event stream; here, we compare this choice with two alternatives: treating randomly selected event signals as keypoints and using Harris corners extracted from RGB images rather than from event signals.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Inspecting the upper row of **Figure 10** reveals that wide intensity swings occur within each frame and that sudden jumps occur across frames, posing a formidable challenge for tracking algorithms and human vision, especially in the third frame. We designate the person visible in the third frame as the tracking target, and the results yielded under different keypoint guidance are gathered in the bottom row. The proposed alignment strategy attains the best outcomes in both restored image quality and tracking accuracy. In contrast, the output from registration driven by random event signals yields only a minor improvement in image quality and results in a less compact bounding box. In contrast, registration steered by RGB-derived corners provides little benefit—an observation that further validates the decision to incorporate event cameras in such harshly lit scenes. The poorer showing of the two baseline implementations mainly arises from their failure to correctly establish temporal correspondences, owing to a scarcity of sufficiently distinctive features.

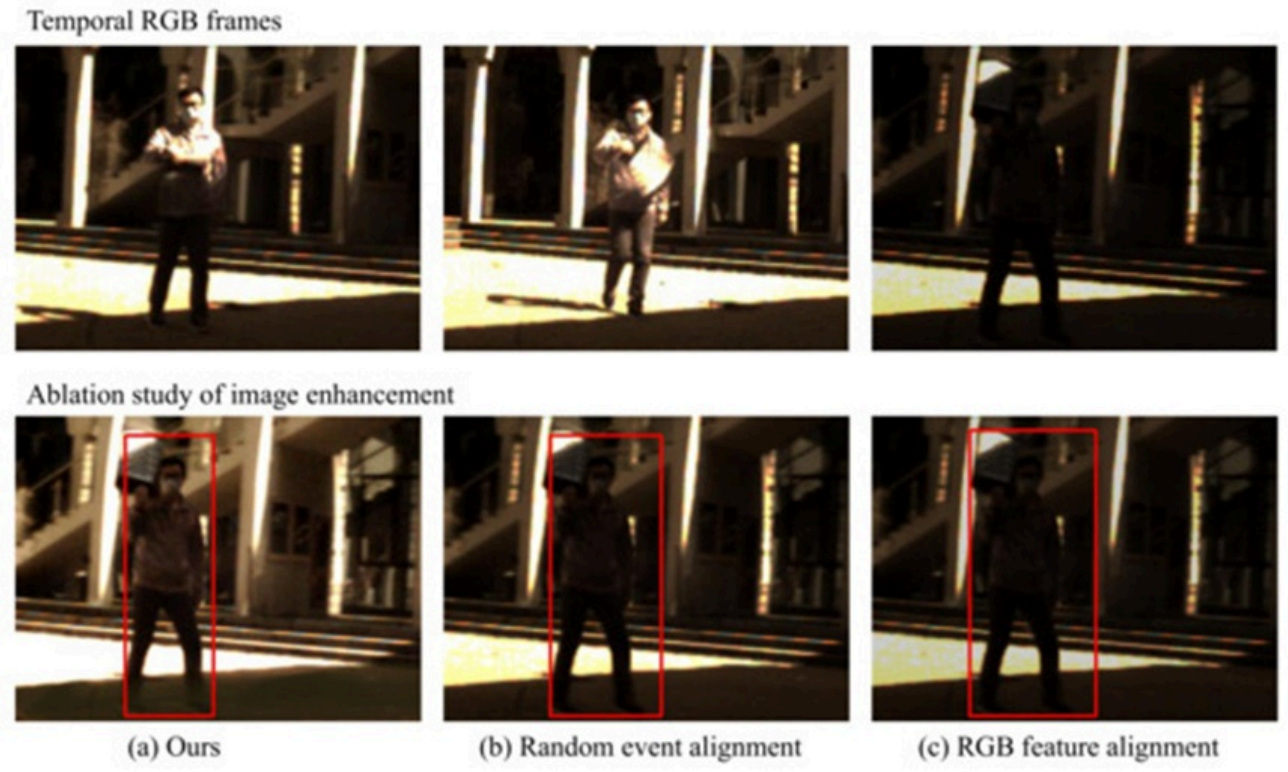


Figure 10. An example illustrating the outcomes of the ablation study. The top row shows the RGB frames from a high-dynamic-range scene. The lower row presents the image enhancement and object-tracking results for three distinct temporal registration guidance methods, with the person in the third frame (the darkest and most challenging) chosen as the target. From left to right: keypoint alignment based on the proposed event-driven Harris corners, random event signals, and Harris corners extracted from RGB frames.

Visible-wavelength optical cameras provide rich descriptive information about the scene, enhancing drone environmental awareness. Yet, harsh illumination creates high-dynamic-range conditions (for instance, at night, near entryways or exit points, etc.) and obstructs reliable scene interpretation. To broaden the practical deployment envelope of visible-wavelength cameras in real-world environments, we propose a dual-sensing architecture that leverages the strengths of event cameras to enhance both the imaging quality delivered by the RGB sensor and downstream object-tracking performance.

The proposed event-assisted robust object tracker leverages two defining traits of event signals: illumination-resilient imaging and rapid temporal response. These favorable and distinctive qualities enable the reconstruction of uninterrupted corner-point paths, which guide the temporal registration of high-dynamic-range RGB frames. Registration serves a

© 2023 The Author(s). | Published by Agro Intelligence Research Press

pivotal function in compensating for intensity fluctuations. Experimentally, the proposed event-assisted robust object tracker performs effectively in high-dynamic-range environments beyond the reach of conventional RGB cameras.

The proposed algorithm surpasses both configurations that take only RGB frames as input and those that directly consume two input modalities, with these benefits holding across a diverse set of applications. From the comparative assessment, we can reach the following two conclusions: (i) Under harsh illumination, the quality of RGB imagery profoundly shapes the success of downstream tasks. To ensure reliable performance in tasks such as object tracking, the RGB frames must first be enhanced. (ii) Event signals, functioning as a lightweight and efficient sensing modality, are capable of capturing critical information in scenes featuring rapid motion. Moreover, event signals remain largely unaffected by lighting conditions and can therefore support scene perception under extreme illumination.

· Limitations. The proposed algorithm has two principal limitations. First, its computational complexity makes direct deployment on a UAV difficult due to the limited onboard processing resources. Achieving UAV-level deployment requires additional network optimization to enable lightweight execution. Second, because the event camera exclusively registers changes in scene intensity, it struggles to detect targets that stay approximately stationary relative to the event camera. Hence, incorporating complementary sensing modalities is necessary to achieve highly resilient object tracking.

· Potential extensions. In subsequent work, we intend to examine event signal characteristics in greater depth and to devise neural architectures better suited to event data, with a focus on lightweight network design and efficient learning. Moreover, we plan to integrate sensing modalities such as LiDAR and IMUs to enable depth-informed 3D scene representations.

Acknowledgements

None

Conflict of interest

Author Heng Luan was employed by TravelSky Technology Ltd.'s Research and Development Center. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Financial support

This work was supported by the Ministry of Science and Technology of China (Grant No. 2020AAA0108202), National Natural Science Foundation of China (grant numbers 61931012, 62171258).

Ethics statement

None

Received: 10 Jul 2022 Revised: 14 Sep 2022 Accepted: 30 Sep 2022
Published online: 25 November 2023

Rights and permissions

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

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. Biol Conserv. 2016;198:60-9.

2. Duffy JP, Cunliffe AM, DeBell L, Sandbrook C, Wich SA, Shutler JD, et al. Location, location, location: Considerations when using lightweight drones in challenging environments. Remote Sens Ecol Conserv. 2018;4:7-19.

3. Zhang Y, He D, Li L, Chen B. A lightweight authentication and key agreement scheme for Internet of Drones. Comput Commun. 2020;154:455-64.

4. McNeal GS. Drones and the future of aerial surveillance. Georg Wash Law Rev. 2016;84:354.

5. Akram MW, Bashir AK, Shamshad S, Saleem MA, AlZubi AA, Chaudhry SA, et al. A secure and lightweight drones-access protocol for smart city surveillance. IEEE Trans Intell Transp Syst. 2021;23:19634-43.

6. Guvenc I, Koohifar F, Singh S, Sichitiu ML, Matolak D. Detection, tracking, and interdiction for amateur drones. IEEE Commun Mag. 2018;56:75-81.

7. Bamburry D. Drones: Designed for product delivery. Des Manag Rev. 2015;26:40-8.

8. Panda SS, Rao MN, Thenkabail PS, Fitzgerald JE. Remote sensing systems—platforms and sensors: Aerial, satellite, UAV, optical, radar, and LiDAR. In: Remotely Sensed Data Characterization, Classification, and Accuracies. Boca Raton (FL): CRC Press; 2015. p. 37-92.

9. Jeong N, Hwang H, Matson ET. Evaluation of low-cost lidar sensor for application in indoor UAV navigation. In: Proc IEEE Sens Appl Symp. Seoul, Republic of Korea; 2018. p. 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 Biol Biochem. 2011;43:1398-410.

11. Chen P, Dang Y, Liang R, Zhu W, He X. Real-time object tracking on a drone with multi-inertial sensing data. IEEE Trans Intell Transp Syst. 2017;19:131-9.

12. Wen L, Zhu P, Du D, Bian X, Ling H, Hu Q, et al. VisDrone-SOT2018: The vision meets drone single-object tracking challenge results. In: Proc Eur Conf Comput Vis Workshops. Munich, Germany; 2018. p. 469-95.

13. Bartak R, Vykovský A. Any object tracking and following by a flying drone. In: Proc Mex Int Conf Artif Intell. Cuernavaca, Mexico; 2015. p. 35-41.

14. Zhang H, Wang G, Lei Z, Hwang JN. Eye in the sky: Drone-based object tracking and 3D localization. In: Proc ACM Int Conf Multimed. Nice, France; 2019. p. 899-907.

15. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH. Fully-convolutional siamese networks for object tracking. In: Proc Eur Conf Comput Vis Workshops. Amsterdam, The Netherlands; 2016. p. 850-65.

16. Danelljan M, Bhat G, Shahbaz Khan F, Felsberg M. ECO: Efficient convolution operators for tracking. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. Honolulu, HI, USA; 2017. p. 6638-46.

17. Dai K, Wang D, Lu H, Sun C, Li J. Visual tracking via adaptive spatially-regularized correlation filters. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. Long Beach, CA, USA; 2019. p. 4670-9.

18. Li P, Chen B, Ouyang W, Wang D, Yang X, Lu H. GradNet: Gradient-guided network for visual object tracking. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. Long Beach, CA, USA; 2019. p. 6162-71.

19. Mitrokhin A, Fermüller C, Parameshwara C, Aloimonos Y. Event-based moving object detection and tracking. In: Proc IEEE/RSJ Int Conf Intell Robots Syst. Madrid, Spain; 2018. p. 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: Proc AAAI Conf Artif Intell. New York, NY, USA; 2020;34:10534-41.

21. Burdziakowski P, Bobkowska K. Lighting conditions—accuracy considerations. Sensors. 2021;21:3531.

22. Wisniewski M, Rana ZA, Petrunin I. Drone model classification using convolutional neural network trained on synthetic data. J Imaging. 2022;8:218.

23. Onzon E, Mannan F, Heide F. Neural auto-exposure for high-dynamic range object detection. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. Nashville, TN, USA; 2021. p. 7710-20.

24. Mahlknecht F, Gehrig D, Nash J, Rockenbauer FM, Morrell B, Delaune J, et al. Exploring event camera-based odometry for planetary robots. IEEE Robot Autom Lett. 2022;7:8651-8.

25. Debevec PE, Malik J. Recovering high dynamic range radiance maps from photographs. In: Seminal Graphics Papers: Pushing the Boundaries. New York, NY, USA: Association for Computing Machinery; 2023;2:643-52.

26. Kang SB, Uyttendaele M, Winder S, Szeliski R. High dynamic range video. ACM Trans Graph. 2003;22:319-25.

27. Mangiat S, Gibson J. High dynamic range video with ghost removal. In: Proceedings of the Applications of Digital Image Processing; San Diego, CA, USA, 1–5 August 2010; National Council of Teachers of Mathematics: Reston, VA, USA, 2010; 7798:307-14.

28. Mangiat S, Gibson J. Spatially adaptive filtering for registration artifact removal in HDR video. In: Proceedings of the IEEE International Conference on Image Processing; Brussels, Belgium, 2011. p. 1317-20.

29. Kalantari NK, Ramamoorthi R. Deep high dynamic range imaging of dynamic scenes. ACM Trans Graph. 2017;36:1-12.

30. Gryaditskaya Y. High Dynamic Range Imaging: Problems of Video Exposure Bracketing, Luminance Calibration and Gloss Editing. Ph.D. Thesis, Saarland University, Saarbrücken, Germany, 2016.

31. Hajisharif S, Kronander J, Unger J. Adaptive dualISO HDR reconstruction. EURASIP J Image Video Process. 2015;2015:41.

32. Heide F, Steinberger M, Tsai YT, Rouf M, Pająk D, Reddy D, et al. Flexisp: A flexible camera image processing framework. ACM Trans Graph. 2014;33:1-13.

33. Cai J, Gu S, Zhang L. Learning a deep single image contrast enhancer from multi-exposure images. IEEE Trans Image Process. 2018;27:2049-62.

34. Nayar SK, Mitsunaga T. High dynamic range imaging: Spatially varying pixel exposures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Hilton Head, SC, USA; 2000; 1:472-9.

35. Zhao H, Shi B, Fernandez-Cull C, Yeung SK, Raskar R. Unbounded high dynamic range photography using a modulo camera. In: Proceedings of the IEEE International Conference on Computational Photography; Houston, TX, USA, 2015. p. 1-10.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

36. Gallego G, Delbrück T, Orchard G, Bartolozzi C, Taba B, Censi A, et al. Event-based vision: A survey. IEEE Trans Pattern Anal Mach Intell. 2020;44:154-80.

37. 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; Nashville, TN, USA; 2021. p. 1403-9.

38. Lagorce X, Meyer C, Ieng SH, Filliat D, Benosman R. Asynchronous event-based multikernel algorithm for high-speed visual features tracking. IEEE Trans Neural Netw Learn Syst. 2014;26:1710-20.

39. Rebecq H, Ranftl R, Koltun V, Scaramuzza D. High speed and high dynamic range video with an event camera. IEEE Trans Pattern Anal Mach Intell. 2019;43:1964-80.

40. 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; Melbourne, Australia; 2014. p. 686-9.

41. Ni Z, Pacoret C, Benosman R, Ieng S, RÉGNIER S. Asynchronous event-based high speed vision for microparticle tracking. J Microsc. 2012;245:236-44.

42. Tulyakov S, Gehrig D, Georgoulis S, Erbach J, Gehrig M, Li Y, et al. Time lens: Event-based video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Nashville, TN, USA; 2021. p. 16155-64.

43. 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; New Orleans, LA, USA; 2022. p. 17755-64.

44. 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; Seattle, WA, USA; 2020. p. 1669-78.

45. Bardow P, Davison AJ, 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; Las Vegas, NV, USA; 2016. p. 884-92.

46. Wan Z, Dai Y, Mao Y. Learning dense and continuous optical flow from an event camera. IEEE Trans Image Process. 2022;31:7237-51.

47. Akolkar H, Ieng SH, Benosman R. Real-time high speed motion prediction using fast aperture-robust event-driven visual flow. IEEE Trans Pattern Anal Mach Intell. 2020;44:361-72.

48. Ramesh B, Zhang S, Lee ZW, Gao Z, Orchard G, Xiang C. Long-term object tracking with a moving event camera. In: Proceedings of the British Machine Vision Conference; Newcastle, UK; 2018; p. 241.

49. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference; Munich, Germany, Switzerland; 2015. p. 234-41.

50. Jung I, Son J, Baek M, Han B. Real-time MDNet. In: Proceedings of the European Conference on Computer Vision; Munich, Germany; 2018. p. 83-98.

51. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); Long Beach, CA, USA; 2017;30.

52. Wang X, Li J, Zhu L, Zhang Z, Chen Z, Li X, et al. VisEvent: Reliable Object Tracking via Collaboration of Frame and Event Flows. arXiv. 2023;arXiv:2108.05015.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

53. 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; Long Beach, CA, USA; 2019 .p. 4277-86.

© 2023 The Author(s). | Published by Agro Intelligence Research Press