RGB-Event Fusion for Moving Object Detection in Autonomous Driving
Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, Dominique Ginhac

To cite this version:

HAL Id: hal-04048853
https://hal.science/hal-04048853
Submitted on 28 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
RGB-Event Fusion for Moving Object Detection in Autonomous Driving

Zhuyun Zhou\textsuperscript{1}, Zongwei Wu\textsuperscript{1,2}, Rémi Boutteau\textsuperscript{3}, Fan Yang\textsuperscript{1}, Cédric Demonceaux\textsuperscript{1,4}, Dominique Ginhac\textsuperscript{1}

Abstract—Moving Object Detection (MOD) is a critical vision task for successfully achieving safe autonomous driving. Despite plausible results of deep learning methods, most existing approaches are only frame-based and may fail to reach reasonable performance when dealing with dynamic traffic participants. Recent advances in sensor technologies, especially the Event camera, can naturally complement the conventional camera approach to better model moving objects. However, event-based works often adopt a pre-defined time window for event representation, and simply integrate it to estimate image intensities from events, neglecting much of the rich temporal information from the available asynchronous events. Therefore, from a new perspective, we propose RENet, a novel RGB-Event fusion Network, that jointly exploits the two complementary modalities to achieve more robust MOD under challenging scenarios for autonomous driving. Specifically, we first design a temporal multi-scale aggregation module to fully leverage event frames from both the RGB exposure time and larger intervals. Then we introduce a bi-directional fusion module to attentively calibrate and fuse multi-modal features. To evaluate the performance of our network, we carefully select and annotate a sub-MOD dataset from the commonly used DSEC dataset. Extensive experiments demonstrate that our proposed method performs significantly better than the state-of-the-art RGB-Event fusion alternatives. The source code and dataset are publicly available at: https://github.com/ZZY-Zhou/RENet.

I. INTRODUCTION

With the development of deep learning methods, researchers have reported great success in modeling the motion behaviors of different traffic participants [1]–[3]. However, most existing methods are frame-based and their performances are highly dependent on video quality, i.e., the frame rate and the image quality [4]–[6]. The former (frame rate) is essential to model the temporal consistency and the motion behavior, while the latter (image quality) influences the accuracy of object detection within a single frame. Therefore, despite significant advances in deep networks, the inherent limits of frame-based cameras are one of the main performance bottlenecks for moving object detection.

Recent advances in bio-inspired event cameras have drawn great research attention for autonomous vehicles [7]–[9]. The rich temporal information provided by events can help to better avoid collisions with moving traffic agents and thus yield safer autonomous driving. In fact, when dealing with complex lighting scenes, such as driving during night or passing through a tunnel, the performance of frame-based methods may degrade severely [10]–[13]. Meanwhile, event cameras are more robust in these adverse visual conditions since they can quickly adapt to light changes thanks to its low latency. Nevertheless, event data often provide wider object boundaries than RGB frames. This is mainly due to the synchronization between RGB and events since RGB frames are obtained during a short exposure time, while event cameras work in an asynchronous mode at a higher temporal resolution. Therefore, how to fully leverage the heterogeneous but mutually complementary modalities is yet an open research topic.

In the literature, a common practice [15], [18]–[20] is to gradually merge event features in the RGB stream to guide the feature modeling as shown in Fig. 1. Despite the achieved plausible results, these works are often unidirectional and cannot thoroughly leverage modality-specific cues during the encoding stages, i.e., RGB-only, event-only, and fused features. Moreover, most existing works choose a pre-defined time interval as a hyper-parameter for event representation, which yields two major shortages: first, the fixed event representation cannot adapt to different scenarios without handcraft adjusting; secondly, a small range of event intervals, such as the exposure time of RGB camera, cannot fully leverage the rich temporal cues from the event...
data, while a long range of event intervals adds significant misaligned response.

To address the aforementioned issues, in this paper, we propose a novel RGB-Event network, named RENet, for moving object detection. Different from previous works basing on single-range events, as shown in Fig. 1, our event stream takes multi-range events as inputs to better benefit from the rich temporal cues. Specifically, for each event representation, we first apply convolutions to project the input event maps to a latent space. We then apply pooling with different kernel sizes to deal with events from different time ranges, i.e., larger pooling windows for longer-range events. Our motivation comes from the observation that the event frames of long-time ranges introduce local motion blur around the object boundaries. The motion blurs are essential for the detection of moving objects, while being not well-aligned with the image frame obtained during the exposure time. Therefore, we apply different sizes of max-pooling to preserve the most informative motion features and improve the scale invariance within a local region, yielding a simple yet efficient way to aggregate rich temporal events in a coarse-to-fine manner.

Once the multi-range events are fused in the latent space, we feed them through the dual-residual networks along with RGB features. To attentively integrate multi-modal features, we introduce a bi-directional calibration module that first improves each input feature by attending to the channel and the spatial dimensions of the input features and then realizes a cross-modal calibration. Finally, the cross-calibrated features are merged and further integrated with the hierarchical features, if any, to form the shared output. To the best of our knowledge, we are the first to apply attention modules for RGB-Event (moving) object detection tasks. Additionally, we propose DSEC-MOD, a new dataset dedicated to moving object detection from the widely used RGB-Event dataset DSEC [21]. DSEC-MOD contains moving objects with automatically labeled and manually checked bounding boxes. Extensive comparisons performed on DSEC-MOD demonstrate the effectiveness of RENet against state-of-the-art RGB-Event fusion alternatives.

To summarize, our main contributions are three-fold:

- We design a multi-range aggregation module to fully leverage the rich temporal information from events, which are crucial for moving object detection.
- We introduce a novel middle fusion design for RGB-Event fusion that models both modality-specific and shared representations.
- We propose a novel DSEC-MOD dataset with automatically labeled and manually verified annotation to encourage the development of moving object detection with events. We benchmark various state-of-the-art (SOTA) fusion modules on our dataset and show that RENet significantly outperforms the fusion alternatives.

II. RELATED WORK

Event Datasets: Unlike traditional datasets for global shutter RGB cameras, the number of available public event datasets is much fewer, and the types of tasks that may be practiced with these datasets are also limited. For example, several datasets [21]–[24] only provide depth information as ground truth, including the two most widely used DSEC [21] and MVSEC [24]. To encourage the development of segmentation and detection tasks for autonomous driving, recent researches propose to induce from these two datasets and create sub-datasets for specific tasks. For example, MVSEC-NIGHT [22] induces from MVSEC and generates car annotations with the help of the pre-trained YOLO-v3 [25] for night scenarios. Sharing a similar motivation, [15] uses the pre-trained YOLO-v5 [26] to create the bounding boxes for objects from the DSEC dataset. There are also several other datasets targeting segmentation tasks, such as the DET dataset [27] which provides annotation for lane detection. However, these works focus more on 2D segmentation, with few works taking the temporal consistency and object motion into account. For autonomous driving, understanding object motions is crucial for driving safety. Therefore, we propose an RGB-Event dataset, called DSEC-MOD, to encourage further research on moving object detection.

Event Processing: Event cameras have recently drawn great attention thanks to their asynchronous characteristic. As bio-inspired sensors, several event-by-event researches aim to develop novel spiking neural network [28]–[30] to process this novel sensor information [31], [32]. However, there are few works dealing with RGB-Event inputs, especially for multi-modal sensor fusion.

From another perspective, events can be naturally considered as 4D inputs \((x, y, p, t)\), where \((x, y)\) stands for the spatial resolution as that of images, \(p\) stands for the polarity, and \(t\) stands for the temporal axis. Until nowadays, there is no conventional event representation for video applications as suggested in previous works [18], [33]. In the literature, there are three popular representations: frame-like methods [34]–[37] that focus more on dealing with \((x, y, p)\) information, time-surface methods [38]–[40] that focus more on \((x, y, t)\) cues, and voxel-based methods [19], [41]–[43] that deal with all dimensions at the same time \((x, y, p, t)\). While voxel-based methods can fully explore both space-time information, its 4D structure requires higher computational cost compared to frame-based and time-surface methods, which are not suitable for real-time applications such as autonomous driving. In this paper, we propose a novel event representation for moving object detection. We build upon frame-based approaches and additionally integrate the time information through a temporal multi-scale aggregation, resulting in a simple yet effective manner to explore spatio-temporal cues.

**RGB-Event Fusion:** With the development of event cameras, RGB-Event multi-modal fusion has drawn increasing research attention. Some works adopt a pre-trained network to realize event-frame conversion. Then, several works directly process the generated video in the same way as dealing with RGB video to achieve the target application. A recent work [44] applies knowledge distillation to guide the learning on videos generated from events. However, we argue that there is a significant difference between knowledge
distillation and multi-modal fusion. Other works take RGB and events from the input side and realize features during the feature modeling. [16], [45] proposes a late-fusion design to integrate events into RGB streaming to guide the feature decoding. Meanwhile, [15] adopts a middle-fusion design through simple concatenation convolution. [18] further leverages the transformer’s attention to better guide the feature fusion. Sharing a similar idea, we also follow the middle-fusion design to fuse RGB and events. Different from [15], [18] with uni-directional fusion design, as shown in Fig. 1, we explicitly model both modality-specific features and shared representation. Furthermore, we introduce a cross-modal calibration module to attentively correct the noisy response in each modality before feature fusion. Different from [18] which focuses on the channel axis, our attention module further attends to the spatial dimension, which plays an important role for object localization within a frame.

III. OUR WORK

Fig. 2 presents the overall framework of our network RENet. For simplicity, we only illustrate the case for single-frame RGB-Event inputs. In practice, inputs use several frames from video as well as the associated events.

Our network is composed of a tailored stemming layer to merge temporal multi-scale events (E-TMA see Sec. III-A), a dual encoder network to extract features, a bi-directional calibration module to firstly improve and then generate the shared representation (BDC see Sec. III-C), and a decoder to output the bounding box of moving objects. Specifically, the multi-scale events are first fed into our proposed E-TMA to fully leverage the rich temporal motion cues. Then, we adopt a dual but discrepant residual streaming design to extract semantic maps from each modality. To realize the multi-modal fusion, we attend to both the channel and spatial dimensions of each modality and then enable a cross-modal calibration. This process aims to neglect inherent single-modal noisy response and improve the feature modeling. Then, the enhanced multi-modal features together with the hierarchical features are fused in a coarse-to-fine manner.

Finally, the encoded features are fed into the MOD detector [46] to produce the bounding box. The details will be introduced in the following sections.

A. E-TMA: Event-based Temporal Multi-scale Aggregation

We observe from Fig. 3 that objects can slightly move from one position to another, i.e., there are more events on object boundaries while the temporal range of events becomes larger. However, the position changes are not significant nor sudden. Inspired by this observation, we design an aggregation module to establish the temporal relationship between multi-range event features \(E_i\). The aggregation must have two properties. First, it must maintain the homogeneous semantic cues across different ranges since the overall environment remains the same. Second, it must also be sensitive to spatially and temporally moving positions. Most existing works often process events within a pre-defined range, neglecting rich dynamic characteristics of event cameras.

To address this issue, we propose a multi-scale event aggregation module. Here, the term Scale does Not stand for the spatial resolution as conventionally used in image-based tasks. Instead, it stands for the different Temporal ranges widely used in the signal processing domain. Mathematically, during a specific temporal pool \(T = \{T_i|i \in N^*\}\) under the rule that \(T_i \leq T_{i+k}\) with \(k \in N\), the groups of events \(e_i\) are generated respectively. In our case, we take three scales of events as inputs, i.e., the events obtained from the RGB frame exposure time \(e_1\), the events obtained from an intermediate scale which is double the exposure time in our case \(e_2\), and the events obtained during one RGB frame time \(e_3\). To deal with multi-scale events, we first project the input events into a latent feature space. The projection module, denoted as \(\eta\), is similar to the first VGG layer which is composed of a combination of Convolution, Batch Normalization, and ReLU activation. Each scale of events is fed into the same projection module to generate the event feature map. In other words, the projection weights are shared for each scale event input. The projection step can be expressed as:

\[
E_i' = \eta(e_i), i \in 1, 2, 3.
\]
To model the objects moving within a local region, pooling operations with different window sizes are applied to the event features: features from long-range events are processed with coarser pooling, while features from small-range events are processed with finer pooling. Formally, let the pooling with different window sizes $pool_i = \{pool_1; pool_2; pool_3| size(pool_1) < size(pool_2) < size(pool_3)\}$, we apply max pooling on top multi-scale event features $E'_i$:

$$e'_i = pool_i(E'_i), i \in 1, 2, 3.$$  \hspace{1cm} (2)

Note that due to the different sizes of the pooling kernels, the pooled features do not share the same resolution. To merge them together, upsampling is applied to smaller-shape features to adjust the resolution. Finally, all features are concatenated together and fed into a convolution to generate the event representation $f_e$. We have:

$$f_e = \text{Conv}(\text{Concat}(e'_1; \text{up}(e'_2); \text{up}(e'_3))),$$  \hspace{1cm} (3)

where $\text{up}$ denotes the upsampling to match the resolution. The aggregated feature is later fed into the residual encoder [47] for semantic modeling.

Our event modeling is significantly different from previous works. Several works simply take a pre-defined range of events as input [15], [16], which cannot fully leverage the rich temporal cues. A recent concurrent work [18] also proposes to deal with multi-scale events. However, they simply accumulate the polarities of multi-scale events from the input side. Despite its demonstrated effectiveness on image deblurring, this method fails to explicitly distinguish the object boundary and the background, yielding a noisy representation for object detection tasks. Differently, we propose to merge the multi-scale events at the latent space with the help of pooling operations. The choice of pooling is mainly based on its spatial scale and rotation invariant property which fits our application extremely well, as discussed previously. We particularly leverage the max pooling operation to extract the most informative cues in the event features. The motivation is based on the intuition that the object boundary should provide a strong response to distinguish the traffic participant from the background. To the best of our knowledge, we are the first to introduce a learnable aggregation module to extract, preserve, and fuse the most informative event features across different temporal scales. We show in Table I the superiority of our proposed event representation approach over our counterparts.

B. Discrepant Two-Streaming Encoder

Most existing multi-model models apply dual encoders for feature extraction. These encoders do not share weight but share the same architecture. However, we observe that RGB and event maps are significantly different, i.e., the former (RGB) provides dense textual information at each pixel, while the latter (event) only provides sparse activation for the pixels with brightness changes. Moreover, through our empirical experiments, as shown in Table I, we observe that event-only streaming yields significantly lower performance compared to RGB-only streaming. This observation motivates us to design a discrepant two-streaming encoder for feature extraction as shown in Fig. 2, i.e., a deeper encoder for RGB processing with ResNet-101 and a shallower encoder for event processing with Resnet-18. In such a way, we distinguish RGB and event modalities that the event stream plays an assist and auxiliary role. This discrepant design can also help to reduce computational costs.

To realize the multi-modal fusion, RGB and event features must share the same size. However, due to the discrepant encoders, the channel size for our RGB and event features are not the same. Therefore, before realizing the feature fusion, we apply a projection on the encoded event features to match the channel size of RGB features.

C. BDC: Bi-Directional Calibration

RGB-Event encoder fusion is a relatively novel research topic with very few papers compared to other multi-modal fusions such as RGB-Depth and RGB-Point cloud. Previous RGB-Event works or simply merge multi-modal features through simple concatenation convolution [15], or attend to the channel dimension with the help of transformer attention [18]. However, these methods do not explicitly tackle the spatial cues which are essential for segmentation/detection tasks. A recent tentative on knowledge transfer [44] has also shown great success for event processing. Nevertheless, we argue that there is a huge difference between multi-modal fusion and knowledge distillation where the latter cannot fully leverage the cross-modal features compared to the former. Therefore, we seek to design an attentive fusion that can leverage the informative cues from both spatial and channel axis during the encoding stage.

For simplicity, we take the fusion at the highest semantic level as an example. Formally, let the RGB feature $f_R \in \mathbb{R}^{C \times h \times w}$ and the projected event feature $f_e \in \mathbb{R}^{C \times h \times w}$. We first apply a transformation module based on $1 \times 1$ convolution for activation. We obtain the activated feature maps $f_r$ and $f_e$ by:

$$f_r = \text{Conv}_{1 \times 1}(f_R); \quad f_e = \text{Conv}_{1 \times 1}(f_E).$$  \hspace{1cm} (4)
Then we attentively fuse RGB and event features in a coarse-to-fine manner. Specifically, we first use a pixel-wise multiplication along with addition to coarsely enhance the most informative features in each modality. Formally, the enhanced features $f'_r$ and $f'_e$ are computed by:

$$f'_r = f_r \otimes f_e + f_r, \quad f'_e = f_r \otimes f_e + f_e.$$  \hfill (5)

Then we refine the features by explicitly attending to the channel and spatial axes in a separate and successive manner. For each axis, we enable a bi-directional calibration, i.e., we learn the features from one modality and apply it to the other. Then, we merge calibrated multi-modal features while maintaining the most informative features. For simplicity, we show the example with channel calibration. Let $CA$ be the channel attention module from [48]. The cross-calibration along the channel axis can be formulated as:

$$f'^{CA}_r = CA(f'_r) \otimes f'_r, \quad f'^{CA}_e = CA(f'_r) \otimes f'_e + f'_e.$$  \hfill (6)

Sharing the same protocol, we further refine the features spatially. Let $SA$ be the spatial attention from [48], we obtain the final enhanced features $f^{enh}_r$ and $f^{enh}_e$ as follows:

$$f^{enh}_r = SA(f'^{CA}_r) \otimes f'^{CA}_r + f'^{CA}_r, \quad f^{enh}_e = SA(f'^{CA}_r) \otimes f'^{CA}_e + f'^{CA}_e.$$  \hfill (7)

These designs aim to find the most truthful and confident channels/pixels within each modality-specific feature through the deep network. Moreover, cross-calibration can selectively calibrate the noisy features while maintaining the informative cues, yielding a simple yet efficient manner to boost the feature modeling with the help of complementary cues. Finally, the calibrated features are attentively merged through a convolution that learns the contribution weights from the most informative components to form the shared output:

$$f = Conv_{3 \times 3}(Concat(f^{enh}_r \otimes f^{enh}_e; max(f^{enh}_r, f^{enh}_e))).$$  \hfill (8)

Further, the fused output is integrated with hierarchical features, i.e., the output from the previous layer (if any).

IV. EXPERIMENTS

A. Our RGB-Event Moving Object Dataset

To the best of our knowledge, DSEC is the largest RGB-Event dataset for autonomous driving. However, DSEC only proposes the depth ground truth, without object bounding box annotation. [15] proposes to generate object annotation with the pre-trained YOLO-V5 network. However, the used dataset is not available. Moreover, the generated annotation takes both statistics and dynamic objects into account. Therefore, to promote the research of moving object detection with events, we introduce a novel DSEC-MOD dataset.

Specifically, we first calibrate the multi-modal input. Following [21], we project the RGB frames to the event-based coordinates with the help of camera matrices. In this process, RGB and event maps have the same field of view and the same resolution. Then, based on the calibrated RGB videos, we manually label the annotation for moving objects. In our dataset, 8 object classes are taken into account: car, truck, bus, train, pedestrian, cyclist, motorcyclist, and others. Since we aim to build a challenging dataset with multiple moving objects in the scene, our sequences contain at least 3 different types of moving objects. In total, our DSEC-MOD dataset contains 16 sequences (13314 frames), with 11 sequences (10495 frames) for training and 5 other sequences (2819 frames) for testing. To the best of our knowledge, our moving object dataset is the largest of its kind.

B. Experimental Setup

Implementation Details: We choose ResNet-101 and ResNet-18 as our backbones. The input frames are resized to $288 \times 288$. We apply classical data augmentations such as photometric transformation, scale, and location jittering. We follow [46] to use adam optimizer with the initial learning rate 5e-4, then decreasing by a factor of 10 at the 10th epochs and 15th epochs. The whole model is trained for 30 epochs.

Evaluation Metrics: We use frame mean average precision (F. mAP) and video mean average precision (V. mAP) for evaluation. Frame mAP evaluates the detection quality in a single frame, while video mAP also considers the linking of bounding boxes from a temporal perspective [54]–[56].

C. Comparison

Comparison with SOTA fusion alternatives: To the best of our knowledge, there are very few works on RGB-Event fusion during the encoder stage. Moreover, we are the first to conduct a study on moving object detection with RGB-Event input. Therefore, to purely and fairly analyze our performance, we replace our proposed modules with the state-of-the-art fusion alternatives from RGB-Event domain: FPN-Fusion [15], EFNet [18]; and from RGB-D domain: SAGate [51], DCF [52], SPNet [53]. We also compare with famous self-attention modules such as SENet [49], ECAnet [50], and CBAM [48]. Specifically, we maintain the same multi-scale event input with our E-TMA (if with event input), same backbone, same decoder, loss, and same training settings as ours. The only difference is in the RGB-Event fusion module. Note that for all listed SOTA alternatives, we apply the attention modules through middle fusion.

As shown in Table I, compared to both self-attention and cross-modal attention works, our RENet outperforms its counterparts by a wide margin, validating our proposed fusion design. Specifically, compared to FPN-Fusion [15] which is based on middle fusion with concat-conv, our fusion is based on attention which can better deal with informative features across different modalities. Different from EFnet [18] which attends only to the channel dimension, our fusion leverages both channel and spatial attention and realizes in addition a cross-calibration to improve the feature modeling.

Comparison under different illumination conditions: To better understand the contribution of event cameras, we analyze the performance gain under different illumination conditions. The detailed comparison can be found in Table II. It can be seen that with the complementary events, our proposed modules can significantly and consistently boost the baseline performance under different lighting conditions.
TABLE I: Comparison with SOTA fusion alternatives. 0.5 and 0.2 stand for the thresholds of the intersection of union.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Fusion Module</th>
<th>Pub. &amp; Year</th>
<th>F. mAP @0.5 (%)</th>
<th>V. mAP @0.2 (%)</th>
<th>V. mAP @0.5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Setting</td>
<td>RGB only</td>
<td>/</td>
<td>30.75</td>
<td>12.82</td>
<td>4.97</td>
</tr>
<tr>
<td></td>
<td>Event only</td>
<td>/</td>
<td>14.66</td>
<td>4.50</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Early Fusion</td>
<td>/</td>
<td>32.27</td>
<td>17.81</td>
<td>5.43</td>
</tr>
<tr>
<td></td>
<td>Late Fusion</td>
<td>/</td>
<td>34.27</td>
<td>19.08</td>
<td>6.41</td>
</tr>
<tr>
<td>Self-Attention</td>
<td>SENet [49]</td>
<td>CVPR'18</td>
<td>29.28</td>
<td>17.51</td>
<td>3.77</td>
</tr>
<tr>
<td></td>
<td>CBAM [48]</td>
<td>ECCV'18</td>
<td>36.22</td>
<td>16.36</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td>ECANet [50]</td>
<td>CVPR'20</td>
<td>34.49</td>
<td>18.79</td>
<td>6.73</td>
</tr>
<tr>
<td>Cross-Attention</td>
<td>SAGate [51]</td>
<td>ECCV'20</td>
<td>33.62</td>
<td>17.80</td>
<td>5.34</td>
</tr>
<tr>
<td></td>
<td>DCF [52]</td>
<td>CVPR'21</td>
<td>32.20</td>
<td>18.84</td>
<td>4.39</td>
</tr>
<tr>
<td></td>
<td>SPNet [53]</td>
<td>ICCV'21</td>
<td>32.70</td>
<td>14.71</td>
<td>6.27</td>
</tr>
<tr>
<td></td>
<td>EFNet [18]</td>
<td>ECCV'22</td>
<td>35.33</td>
<td>18.05</td>
<td>6.30</td>
</tr>
<tr>
<td></td>
<td>RENet</td>
<td>(Ours)</td>
<td>38.38</td>
<td>19.60</td>
<td>7.06</td>
</tr>
</tbody>
</table>

TABLE II: Empirical comparison under overall and specific illumination conditions.

<table>
<thead>
<tr>
<th>Illum.</th>
<th>Comp.</th>
<th>F. mAP@0.5</th>
<th>V. mAP@0.2</th>
<th>V. mAP@0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGB</td>
<td>38.38 (+7.63)</td>
<td>19.60 (+6.78)</td>
</tr>
<tr>
<td></td>
<td>+ Ours</td>
<td>RGB</td>
<td>30.75</td>
<td>12.82</td>
</tr>
<tr>
<td></td>
<td>+ Ours</td>
<td>RGB-Event</td>
<td>32.28</td>
<td>14.97</td>
</tr>
<tr>
<td>Day</td>
<td></td>
<td>Day</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGB</td>
<td>42.06 (+3.50)</td>
<td>16.81 (+0.03)</td>
</tr>
<tr>
<td></td>
<td>+ Ours</td>
<td>RGB-Event</td>
<td>45.58</td>
<td>16.84</td>
</tr>
<tr>
<td>Twilight</td>
<td></td>
<td>Twilight</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGB</td>
<td>36.02 (+11.58)</td>
<td>25.82 (+12.13)</td>
</tr>
<tr>
<td></td>
<td>+ Ours</td>
<td>RGB-Event</td>
<td>24.44</td>
<td>13.69</td>
</tr>
<tr>
<td>Night</td>
<td></td>
<td>Night</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGB</td>
<td>17.44</td>
<td>15.79</td>
</tr>
<tr>
<td></td>
<td>+ Ours</td>
<td>RGB-Event</td>
<td>25.54 (+8.10)</td>
<td>18.76 (+2.97)</td>
</tr>
</tbody>
</table>

**Fig. 4: Qualitative comparison.** Our network generates more accurate detection with fewer failure cases. Best zoomed in.

**Qualitative Comparison:** Fig. 4 depicts the qualitative comparison between our method and the RGB baseline. We can visualize the significant improvement in detecting objects with the help of events, especially in detecting small objects on the scene that are extremely challenging for RGB baseline. Moreover, when RGB baseline results in false positive due to the inferior lighting condition, our RGB-Event network can reason about more robust detection.

**D. Ablation Studies**

We also conduct ablation studies to analyze the contribution of each component. Specifically, we gradually add our proposed modules on top of the RGB baseline. It can be seen that each proposed method is helpful and can enable performance gain. We also replace our proposed multi-scale event aggregation with the recently proposed multi-scale accumulation [18]. The main difference is that our methods aggregate multi-scale events at the semantic level in a coarse-to-fine manner, while [18] simply concatenates multiple event frames from the input side. Empirically, we can see that the performance drops significantly, which demonstrates the superiority of our proposed module in modeling event data for (moving) object detection.

**V. CONCLUSION**

In this paper, we proposed a novel fusion architecture for moving object detection with RGB-Event inputs. Different from previous works based on single-scale event inputs, we introduce a multi-scale events aggregation to explicitly leverage the rich temporal cues from the asynchronous event sensor. Furthermore, we gradually fuse the heterogeneous RGB and event features in a coarse-to-fine manner, yielding a simple yet efficient cross-calibration mechanism with the help of different forms of attention. Finally, we propose a novel DSEC-MOD dataset to encourage further research on RGB-Event fusion for moving object detection. Extensive comparisons of our proposed dataset validate the effectiveness and robustness of our approach compared to the state-of-the-art fusion alternatives. In our future work, we will test our methods on more recent approaches. We will also collect a real-world RGB-Event dataset for autonomous driving.

**REFERENCES**

object detection network for autonomous driving,” in International Conference on Intelligent Transportation Systems (ITSC), 2018.


