Fusing Event-based and RGB camera for Robust Object Detection in Adverse Conditions

Abhishek Tomy, Anshul Paigwar, Khushdeep Singh Mann, Alessandro Renzaglia, Christian Laugier

Fusing Event-based and RGB camera for Robust Object Detection in Adverse Conditions

Abhishek Tomy, Anshul Paigwar*, Khushdeep S. Mann*, Alessandro Renzaglia, Christian Laugier

***Abstract*— The ability to detect objects, under image corruptions and different weather conditions is vital for deep learning models especially when applied to real-world applications such as autonomous driving. Traditional RGB-based detection fails under these conditions and it is thus important to design a sensor suite that is redundant to failures of the primary frame-based detection. Event-based cameras can complement frame-based cameras in low-light conditions and high dynamic range scenarios that an autonomous vehicle can encounter during navigation. Accordingly, we propose a redundant sensor fusion model of event-based and frame-based cameras that is robust to common image corruptions. The method utilizes a voxel grid representation for events as input and proposes a two-parallel feature extractor network for frames and events. Our sensor fusion approach is more robust to corruptions by over 30% compared to only frame-based detections and outperforms the only event-based detection. The model is trained and evaluated on the publicly released DSEC dataset.**

I. Introduction

A neuromorphic or event-based camera is a bio-inspired sensor that detects the changes in intensity at the pixel level. Event-based cameras rely on Contrast Detector (CD) to detect the changes in the intensity of incoming light at each pixel and when that exceeds a predefined threshold, an event signal is recorded and transmitted asynchronously. Event-based cameras are particularly suited for low-light conditions, such as night-time driving or situations involving high dynamic ranges, and can track motions at high speed and temporal resolution.

Many advances in the deep-learning approaches and applications with frame-based tasks have also been explored using an event-based vision system. Event cameras have been used for object detections [1], visual odometry [2], optical flow estimation [3] and depth prediction [4]. Though event-based cameras have proven their edge in adverse conditions, RGB cameras still perform better in normal conditions.

Considering the domain of autonomous driving which primarily uses frame-based cameras, it becomes crucial to cope with varying weather conditions that limit the applicability of state-of-the-art algorithms. These conditions include snow, fog, frost, and others visualized in Figure 1. Current object-detection models for autonomous driving lack the robustness to perform well in varying conditions [5]. While certain conditions have been modelled including snow [6], fog [7], rain [8], daytime and night-time transitions [5], [9], it is not possible to include all potential environmental conditions. Currently, LiDARs and RADARs are used to complement frame-based cameras to attain robustness to varying illumination and weather conditions [10], [11]. LiDARs however are in general extremely costly and bulky for a commercial autonomous vehicle. Event-based cameras seem a promising alternative for low-cost applications like autonomous delivery robots.

In this work, we propose to fuse the information from an event-based and a frame-based camera for object detection with the goal of obtaining a detection accuracy comparable to RGB detection in normal condition but a more robust model in presence of image corruptions and variations. Accordingly, we evaluate the performance of various fusion models over 15 types of corruptions metrics adopted from [12] that were not part of training. This is a useful performance approximation metric that can be utilized to model natural and internal distortions that a camera may experience.

We perform training and evaluation on the DSEC dataset [13] that contains data from event cameras in stereo setup and is recorded from a dynamic ego vehicle through multiple cities in Switzerland.

The key contributions of this study can be summarized as follows:

- We present a novel sensor fusion method with Event-based and Frame-based camera for improving object detection under adverse conditions. The proposed method uses Feature Pyramid Network to combine information at multiple scales.
- We publicly release an extension of DSEC dataset with ground truth labels to facilitate object detection research using both Event-based and Frame-based camera. Our extension includes 131965 bounding box annotations for Car, Pedestrian and Large Vehicle object categories.
- We show in detail the robustness of the proposed sensor-fusion model against common corruptions in the frame-based camera at various severity level. We perform ablation studies and make comparisons with only RGB/Event based detections to prove the effectiveness of our method.

II. Related Work

A. Event-based detection

Researchers have explored multiple ways to represent the sparse and asynchronous events into a dense tensor that can be fed to a learning based approach. For object detection task, Zhu *et al.* proposed an event-volume representation in which the events are accumulated by weighted interpolation into a discretized time domain [4]. Rebecq *et al.* proposed a state-of-art method to reconstruct a gray-scale image from event

* Authors have equal contribution.

All authors are with University of Grenoble Alpes, Inria, 38000, Grenoble, France; e-mail: firstname.lastname@inria.fr

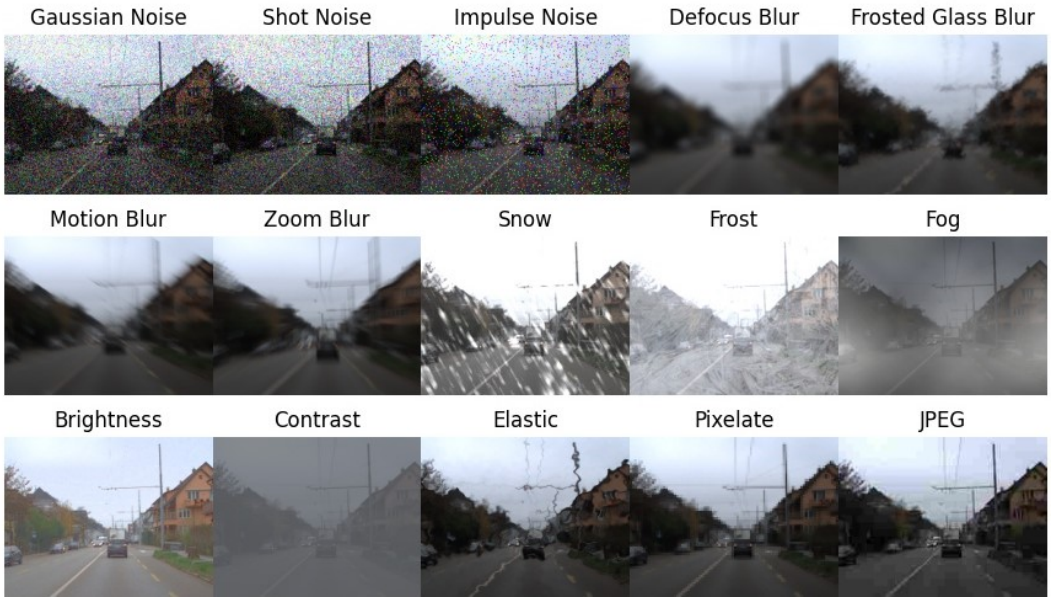

Fig. 1. Visualization of the 15 common corruption types adapted from [14] and applied on a sample image from DSEC dataset. The images correspond to the severity level 3 for each corruption type.

volumes [15]. The reconstructed gray-scale image when applied to standard frame-based object detection networks have shown promising results. Some of the previous works have tried to exploit the sparse nature of events by utilizing spiking neural networks to gain efficiency in processing and handling of events [16]. A Conv-LSTM based state-of-art object detection method that used the spatio-temporal information to improve overall detection was proposed in [1]. In our work, we exploit only individual frame for detection. The ablation studies performed on various input representations show that the event volumes have the best performance for the object detection task [1].

B. Sensor fusion

Fusion of RADAR, camera and LiDAR sensor have been extensively explored for various perception tasks but sensor fusion using event-based camera is relatively a nascent field [17] [18]. Gehrig *et al.* proposed a generalized model of RNNs to handle asynchronous nature of event-based cameras for fusion with other sensor modalities [19]. A general pre-processing step for sensor fusion approaches is to align sensor modalities to a common reference frame. In [20], the radar data is projected to the image plane before concatenating it to the image channels for input to an early sensor fusion model. In [21], the inputs from radar and images are fused at multiple levels, allowing the network to identify at which scale to combine the two modalities. We establish our event-based sensor fusion models inspired from these approaches.

C. Robustness

Numerous studies illustrate that the performance of Convolutional neural networks (CNN) degrade when subjected to image corruptions [12]. Vulnerability of four state-of-art CNN models to image quality distortions, particularly to noise and blur is well demonstrated in [22]. Changing weather conditions impose severe challenges for perception in autonomous driving. Several attempts have been made to model these conditions [5], [6], [9], While it is not possible to model all potential conditions, [14] provide a benchmark for evaluating model performance against common corruptions that are close to natural distortions. A different robustness approach that evaluates the fusion network of RGB camera and Lidar to adversarial attacks have also been undertaken in [23][24].

III. Proposed Method

This section provides a detailed description of our approach starting with the choice of event representation in subsection III-A and in subsection III-B we provide the details to transform RGB images to the viewpoint of events-frames. In subsection III-C, we elaborate on the proposed sensor fusion approach to combine events and RGB images for object detection. Finally, in subsection III-D we provide the adopted image corruption metrics to evaluate robustness.

A. Input Event Representation

An event is a positive or negative signal corresponding to each pixel that is exposed to a change in brightness over a certain threshold value. An event $e_i = (x_i, y_i, t_i, p_i)$ contains the pixel location (x_i, y_i) and the time t_i at which the event is triggered. The polarity of the event ($p_i = \pm 1$) denotes the direction of change.

In this work, we have utilized the voxel grid representation proposed in [4]. The events within a time window ΔT are converted into a $B \times H \times W$ voxel grid where H and W are

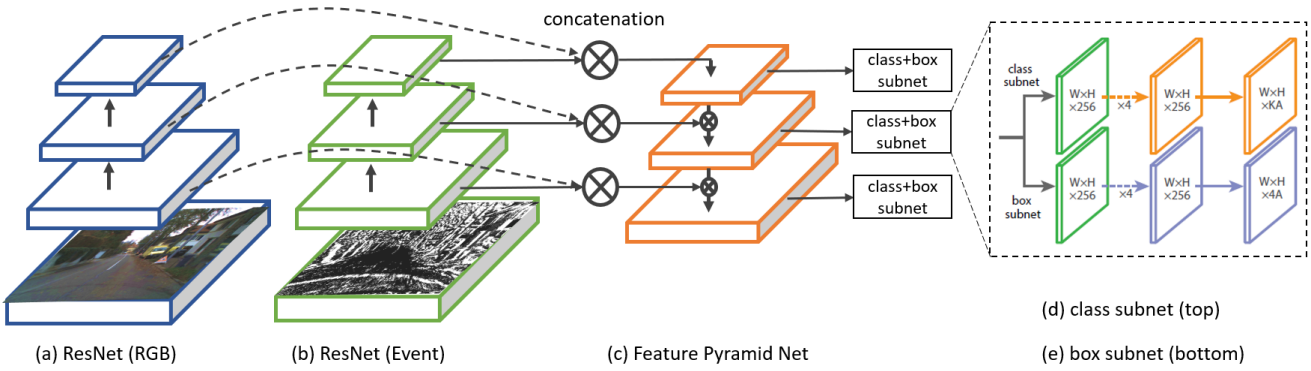

Fig. 2. Network architecture of proposed feature pyramid sensor-fusion model. Event frame and RGB images are passed through a backbone network (ResNet-50) for feature extraction. Pyramidal event and RGB features at the same scale are concatenated before being fed to the feature pyramid network of the RetinaNet-50.

the height and width of the event-frame and B is the number of temporal bins.

$$V(x,y,t) = \sum_i p_i \delta_b(x - x_i)\delta_b(y - y_i)\delta_b(t - t_i^*) \quad (1)$$

where,

$$t_i^* = \frac{B-1}{\Delta T}(t_i - t_1) \quad (2)$$

$$\delta_b(a) = max(0, 1 - |a|) \quad (3)$$

Our models use a time window $\Delta T = 50ms$ and $B = 5$ temporal bins.

Recently, the authors of [15] proposed a method to convert events to a high-quality gray-scale image. The proposed conversion utilizes the voxel grid representation as input to the model. However, the computation of a gray-scale image adds a further pre-processing step that is computationally expensive. We evaluate our object detection models using both the voxel and gray-scale representation in section VI.

B. Homographic transformation

In the DSEC dataset, the events are recorded by a monochrome camera of resolution 640×480 and an RGB camera recording at a 1440×1080 resolution. The baseline of 4.5 cm between two cameras allows for the transformation to a common viewpoint which can be exploited in sensor fusion approaches. In our case, we assume that the imaged scene appears far away from the camera. Since the baseline of the two cameras is small compared to the distances of the scene objects, we utilize the homographic transformation induced by the pure rotation as derived in equation (4) to transform the scene in RGB frame to event-camera frame.

$$P_{event,rgb} = K_{event} * R_{rgb} * R_{event,rgb} * R_{event}^T * K_{rgb}^{-1} \quad (4)$$

where K_{rgb} and K_{event} is the respective intrinsic camera matrix, R_{rgb} and R_{event} is the rotation matrix between distorted and undistorted frame of each cameras, and $R_{event,rgb}$ is the rotation matrix between the RGB camera coordinates system to the event camera.

C. Sensor fusion network: Event + RGB

Our model builds on RetinaNet [25] with a ResNet-50 backbone [26]. The RetinaNet architecture is composed of a backbone feature extractor responsible for computing a feature map over the image, which is passed to a Feature Pyramid Network (FPN), finally, two subnetworks are used for classification and bounding box regression. The FPN has a top-down pathway with 3 pyramidal feature layers going from a coarse feature map to finer maps, along with lateral connection from the backbone network. In our model, RetinaNet is modified to take input from two modalities through two independent streams of feature extractors made up of ResNet-50 for events and RGB frames. Features from both are combined at multiple levels before being fed to the FPN network as shown in Figure 2. This allows the model to extract relevant information from two modalities at different resolutions.

D. Image corruption

All the models presented in this work are trained using clean data, while during testing only camera RGB images were subjected to 15 different corruption types. The performance is evaluated over 5 severity levels for each corruption type. We undertake the image corruption metrics from [12] that was originally introduced in [14]. These corruptions are broadly categorized into four groups as noise, blur, weather, and digital. An illustration of these corruption types under severity level 3 is shown in Figure 1.

E. Robustness metrics

The performance of models for object detection is evaluated using the COCO mAP metric that averages over IoUs between 50% and 95% [27]. Over the corrupted data, the performance is measured as mean performance under corruption (mPC) which is the average mAP over various corruption types and severity levels as shown below:

$$mPC = \frac{1}{N_c}\sum_{c=1}^{N_c}\frac{1}{N_s}\sum_{s=1}^{N_s} mAP_{c,s} \quad (5)$$

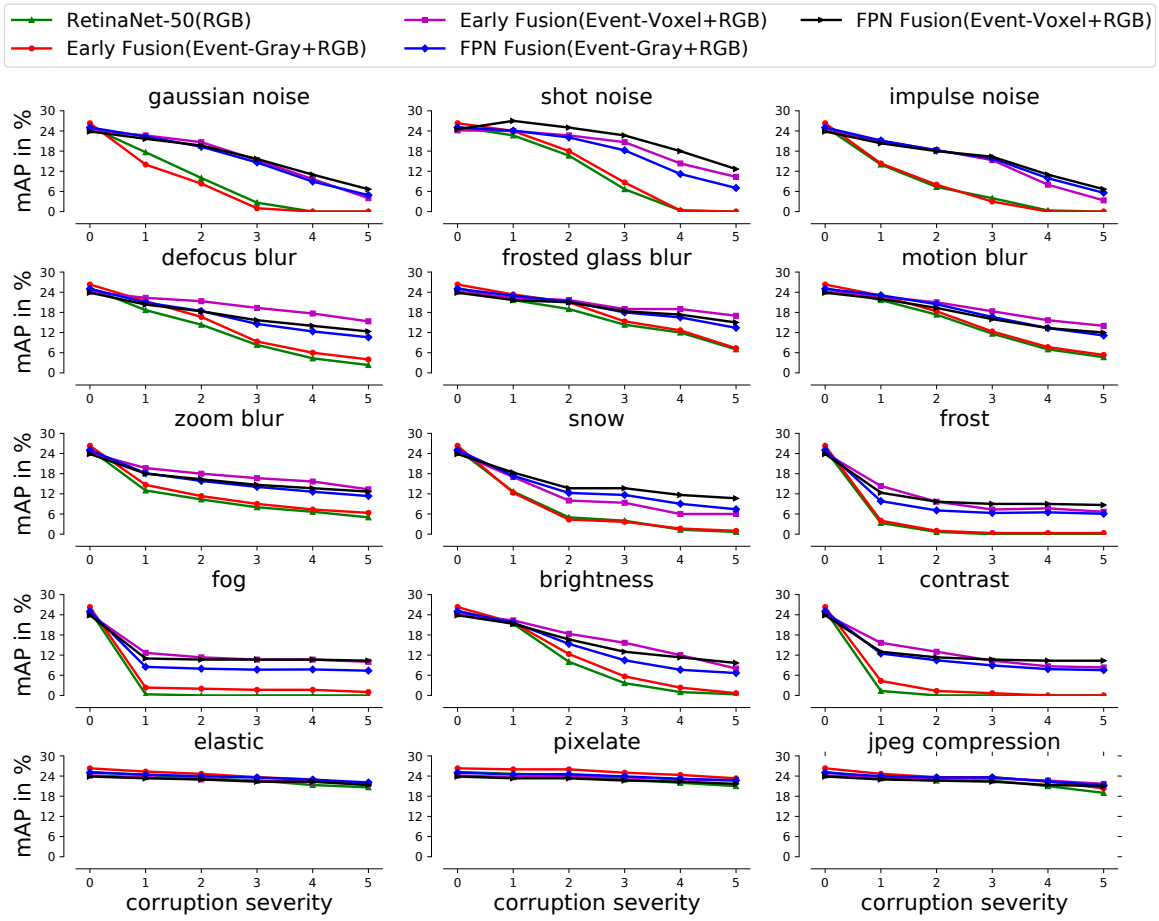

Fig. 3. Model performance mAP (%) subjected to 15 corruption types and 5 severity levels. Severity level 0 implies clean data. We can observe that FPN fusion models are more robust compared to RetinaNet-50 (RGB) and Early fusion (Event-Gray+RGB), particularly for snow, frost, fog, and contrast conditions.

TABLE I

MODEL PERFORMANCE UNDER DIFFERENT CORRUPTION TYPES. FOR EACH CORRUPTION TYPE, RPC (%) IS CALCULATED FOR ALL SEVERITY LEVELS.

		Noise rPC(%)			Blur rPC(%)				Weather rPC(%)				Digital rPC(%)			
Network	mPC(%)	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Fog	Snow	Frost	Bright	Contrast	Elastic	Pixel	JPEG
ReinaNet-50 (RGB)	9.8	24.0	36.7	20.3	38.0	58.7	49.4	34.1	0.3	18.8	3.2	28.8	1.1	89.0	91.1	86.1
Early-fusion (Gray)	10.6	17.7	38.8	19.3	43.6	60.6	50.7	37.0	6.6	17.5	4.6	32.4	4.8	89.7	94.8	87.2
FPN-fusion (Gray)	15.2	56.01	66.02	56.51	61.38	73.56	67.62	57.52	31.34	46.13	28.58	49.37	37.74	93.48	94.97	91.77
Early-fusion (Voxel)	16.0	59.9	74.2	54.6	74.2	81.9	70.3	66.5	44.1	40.0	37.8	59.3	44.7	93.8	96.6	94.9
FPN-fusion events (Voxel)	**16.4**	62.6	97.5	60.6	67.6	78.2	69.3	63.1	44.7	57.6	40.8	60.3	46.6	94.1	95.0	92.5

where $\text{mAP}_{c,s}$ is the performance measure evaluated on corruption type c under severity level s, N_c is the number of corruption types, and N_s is the number of severity levels considered. Further, we use the relative performance under corruption (rPC) metric to measure the relative performance degradation under corruption:

$$rPC = \frac{mPC}{mAP_{clean}} \tag{6}$$

where mAP_{clean} denotes the performance over a clean dataset.

IV. DATASET

Among the available event-based dataset, 1 Mega-pixel event-dataset is one of the largest event-based autonomous driving datasets with over 25M bounding boxes for object detection [1]. RGB images from a frame-based camera were used to generate annotations and then projected to an event-based camera but RGB data is not publicly released.

The DSEC dataset contains data from two event and frame-based cameras in stereo setup and data from both sensor modalities are publicly released [13]. This dataset was recorded from a moving vehicle in challenging illumination conditions and it is particularly suitable to evaluate

TABLE II

OBJECT ANNOTATIONS IN THE DSEC TRAINING DATASET

Categories	***Car***	***Pedestrian***	***Large vehicle (Bus & Truck)***	**Total**
Count	100068	17126	14771	131965
Percentage	0.76	0.13	0.11	1

TABLE III

ABLATION STUDY, MAP ACCURACY WITHOUT THE HOMOGRAPHIC TRANSFORMATION OF RGB IMAGE TO THE EVENT FRAME.

Input	**Model**	***mAP***
Event-Voxel + RGB	Early-fusion	0.14
Event-Voxel + RGB	FPN-fusion	0.19

event-based sensor fusion approaches. However, the dataset lacks annotations for the object detection task. Alternatively, application-specific large-scale event-camera datasets can be generated through simulation for development and experimentation [28]. But this simulation-based dataset poses additional challenges while performing sim2real transfer.

To evaluate our object detection models, we decided to use the DSEC training dataset and generate annotations with an automated labeling protocol similar to [1]. The RGB images are labeled using YOLOv5 from [29]. The labels and bounding box from RGB images are transferred to event-frame using equation (4). The asynchronous event data is labelled at frames corresponding to the timestamps of the RGB images. Bounding boxes of objects within the 640×480 resolution of event-camera are considered. Since the event camera has a lower resolution, bounding boxes with a diagonal smaller than 30 pixels in the event frame are also filtered out. Three dominant object categories were chosen as shown in Table II, to create the final training dataset. The labelled data can be found here [30].

V. TRAINING AND EXPERIMENTS

Our model is trained only on the DSEC dataset and is implemented in PyTorch. The model takes as input information from the left frame-based and event cameras. Training is done to minimize the focal loss between the output and ground truth. The model is trained by using the Adam optimizer [31] with an initial learning rate of $1e^{-4}$ and a batch size of 64. Out of the 41 sequences in the dataset, 30 sequences are considered for training, 3 sequences for validation, and 8 sequences for testing. The training was done on Nvidia GTX 1080 Ti. The implementation of the model is made available in [30].

During the training of the sensor fusion model, the input from the RGB image is made completely blank (zeros) with a probability of 0.15. This forces the sensor fusion model to learn information from the second modality (Event camera). Also, this ensures that the model is robust to failure or corruption from the frame-based camera [21].

A. Ablation study

We study the effectiveness of the proposed homographic transformation of the RGB images as discussed earlier in subsection III-B. Table III shows the results obtained by training our sensor fusion models using original RGB images that have a viewpoint and resolution that is different from the event frame. Comparing these results in Table IV, we observe that with the homographic transformation of RGB images and Event-Voxel, the model achieves an mAP of 0.24 and without the homographic transformation, it drops to 0.19. Also, testing the FPN-fusion model by feeding in an entirely black RGB image and Event-Voxel as input, the model achieves an mAP of 0.12 which is same as the mAP of the event-only models.

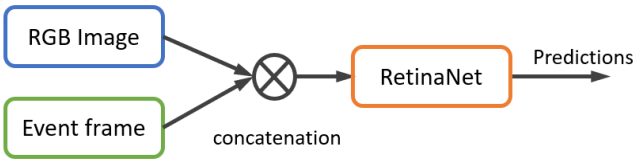

Fig. 4. **Early Fusion:** In this network the RGB and the event-voxels are concatenated before being fed to the RetinaNet.

B. Early sensor fusion network

To evaluate the performance of our proposed FPN based sensor fusion method, we compare it with the performance of a simple early-fusion method. The input from the camera (RGB) and events (Voxel/Gray) are concatenated to create a common input grid and fed as input to a RetinaNet as shown in Figure 4. Previous literature studies have discussed that the early-fusion models are susceptible to perturbations and corruptions of one of the modalities [32]. In this work, we evaluate the model performance and also the capability of an early-fusion method to handle corruption in one of the modalities as discussed further in section VI.

VI. RESULTS

For a thorough comparison, we evaluate models with 7 combinations of input sensor modalities and representations. We used two types of input representations for events as explained in subsection III-A: Event volume representation, indicated by ***Event-Voxel*** and reconstructed gray-scale image from events, indicated by ***Event-Gray***. The input from the frame-based camera is indicated by ***RGB***. For all models including RetinaNet-50 (RGB), the RGB images are transformed to event-frame as explained in subsection III-B. Input from a single modality is evaluated by a RetinaNet with ResNet-50 backbone and all of the results from this model would be indicated by ***RetinaNet-50***. Apart from this, the proposed sensor fusion approach of our work is indicated by ***FPN-fusion*** that captures the essence that the fusion is done at the feature pyramid network and ***Early-fusion*** to indicate the baseline sensor fusion approach from subsection V-B.

The models are primarily evaluated on two metrics: COCO 'mAP' [27] to evaluate the object detection accuracy and 'rPC' to understand relative performance degradation under multiple RGB image corruptions.

A. Model evaluations

The Table IV shows the comparison of various object detection models discussed in this work. The RetinaNet-50 (Event-Voxel/Gray) models have the least mAP and the Early-fusion and FPN-fusion models significantly improve upon this baseline.

Since the labeling of the dataset was done using a frame-based detector, we expect RetinaNet-50 (RGB) to have the best performance, and accordingly, all fusion models achieved a comparable mAP as expected. The Early-fusion (Event-Gray+RGB) model slightly surpasses the RetinaNet-50 (RGB) model. However, it has the second-worst performance on the rPC metric, indicating that the primary information for detection is extracted from RGB and hence is susceptible to image corruptions.

The proposed FPN-fusion model (Event-Voxel+RGB) shows high robustness with 68.7% rPC which is 30% better than the RetinaNet-50 (RGB) model. In general, the models having Event-Voxel input with the raw event data were able to attain a higher rPC in comparison to their counterpart Gray-scale events. This is true for both Early-fusion and FPN-fusion models.

Early-fusion models have shown that with a suitable representation (voxel in this case) of the input, it can achieve higher robustness. On the other hand, ablation study performed in subsection V-A and the results from Table III show that when the two sensor modalities and representation are not in the same viewpoint, the Early fusion model fails to correlate this two information and has a poor performance that is close to RetinaNet-50 (Event-Voxel/Gray) models.

B. Discussion on robustness

Table I provides average rPC(%) over five severity levels. A brief look at the four categories of corruption shows that the RetinaNet-50 (RGB) model is severely affected by noise and weather conditions. Meanwhile, the sensor fusion models, excluding Early-fusion (Event-Gray+RGB), are comparatively robust to all categories. Notably, the weather conditions such as snow, frost, and fog also affect the event camera. Hence, the results do not illustrate the performance of fusion models under weather conditions, but rather an elaborate study on the performance of models with corruptions on the RGB images. However, an event camera has the intrinsic ability to work well in low-light conditions or with high dynamic range scenarios. As seen in Table I, under *Brightness* and *Contrast* corruption type the RetinaNet-50 (RGB) model degrades rapidly compared to the proposed FPN-fusion models.

Figure 3 provides the performance of different models across 15 corruption types and 5 severity levels. Under the *Digital* corruption category, all models including RetinaNet-50 (RGB) model are arguably robust (excluding the *contrast* corruption). Moreover, under severity level 5 (worst corruption scenario), FPN-fusion (Event-Voxel/Gray+RGB) and Early-fusion (Event-Voxel+RGB) models have an mAP close to RetinaNet-50 (Event-Voxel/Gray). This indicates that in case of faulty frame-based sensor or corrupted images, the proposed model is able to fall back on events for object detection.

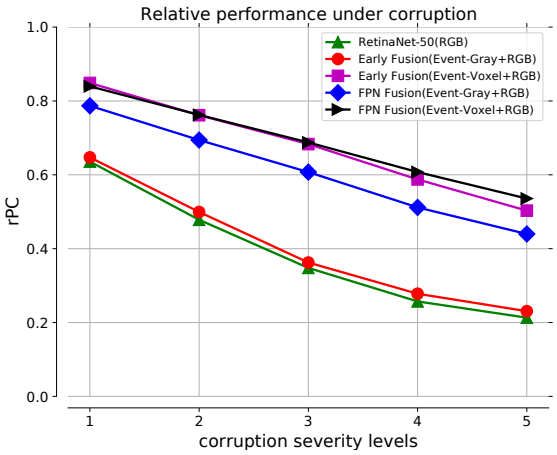

Fig. 5. Relative performance under various severity levels.

TABLE IV

COMPARING THE PERFORMANCE OF THE DIFFERENT PROPOSED MODELS ON MAP AND RPC METRICS.

Input	Model	*mAP*	*rPC (%)*
Event-Voxel	RetinaNet-50	0.12	-
Event-Gray	RetinaNet-50	0.12	-
RGB	RetinaNet-50	0.25	38.6
Event-Gray + RGB	Early-fusion	**0.26**	40.4
Event-Gray + RGB	FPN-fusion	0.25	60.8
Event-Voxel + RGB	Early-fusion	0.24	66.2
Event-Voxel + RGB	FPN-fusion	0.24	**68.7**

CONCLUSIONS AND FUTURE WORK

In this work, we presented the first sensor fusion approach using Event-based and RGB cameras for robust object detection. Our approach combines features from two independent feature extractors corresponding to events and RGB in the feature pyramidal network of RetinaNet. We analyzed the robustness under common image corruption. We demonstrate that our sensor fusion approach has obtained higher robustness for a similar object detection performance. Along with this, to promote further research we have publicly released the labels for object detection in DSEC dataset.

In the future, we would refine our object labels in DSEC dataset through manual annotation and add more object categories. Apart from RetinaNet, various other baseline models can be explored for sensor fusion to achieve a higher mAP and robustness.

ACKNOWLEDGEMENT

This work was conducted within the scope of the ES3CAP (Embedded Smart Safe Secure Computing Autonomous Platform) project.

REFERENCES

[1] E. Perot, P. de Tournemire, D. Nitti, J. Masci, and A. Sironi, "Learning to detect objects with a 1 megapixel event camera," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[2] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.

[3] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Ev-flownet: Self-supervised optical flow estimation for event-based cameras," *Robotics: Science and Systems (RSS)*, 2018.

[4] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 989–997.

[5] D. Dai and L. Van Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3819–3824.

[6] A. Von Bernuth, G. Volk, and O. Bringmann, "Simulating photo-realistic snow and fog on existing images for enhanced cnn training and evaluation," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 41–46.

[7] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[8] D. Hospach, S. Mueller, W. Rosenstiel, and O. Bringmann, "Simulation of falling rain for robustness testing of video-based surround sensing systems," in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2016, pp. 233–236.

[9] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7374–7383.

[10] H. Rashed, M. Ramzy, V. Vaquero, A. El Sallab, G. Sistu, and S. Yogamani, "Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[11] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 682–11 692.

[12] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *ArXiv*, vol. abs/1907.07484, 2019.

[13] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021.

[14] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[15] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[16] N. Kasabov, K. Dhoble, N. Nuntalid, and G. Indiveri, "Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition," *Neural Networks*, vol. 41, pp. 188–201, 2013.

[17] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[18] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7345–7353.

[19] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2822–2829, 2021.

[20] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8311–8317.

[21] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE, 2019, pp. 1–7.

[22] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *8th International Conference on Quality of Multimedia Experience, QoMEX 2016*. Institute of Electrical and Electronics Engineers Inc., 2016, p. 7498955.

[23] J. Tu, H. Li, X. Yan, M. Ren, Y. Chen, M. Liang, E. Bitar, E. Yumer, and R. Urtasun, "Exploring adversarial robustness of multi-sensor perception systems in self driving," *arXiv preprint arXiv:2101.06784*, 2021.

[24] M. Abdelfattah, K. Yuan, Z. J. Wang, and R. Ward, "Adversarial attacks on camera-lidar models for 3d car detection," *arXiv preprint arXiv:2103.09448*, 2021.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[28] H. Rebecq, D. Gehrig, and D. Scaramuzza, "Esim: an open event camera simulator," in *Conference on Robot Learning*. PMLR, 2018, pp. 969–982.

[29] Ultralytics. Yolov5. [Online]. Available: https://github.com/ultralytics/yolov5

[30] FPN-Fusion. Event+rgb. [Online]. Available: https://github.com/abhishek1411/event-rgb-fusion

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] S. Wang, T. Wu, and Y. Vorobeychik, "Towards robust sensor fusion in visual perception," *arXiv preprint arXiv:2006.13192*, 2020.