AFOD: An Adaptable Framework for Object Detection in Event-based Vision

Shixiong Zhang, Wenmin Wang, *Member, IEEE*, Honglei Li

Abstract—Event-based vision is a novel bio-inspired vision that has attracted the interest of many researchers. As a neuromorphic vision, the sensor is different from the traditional frame-based cameras. It has such advantages that conventional frame-based cameras can't match, e.g., high temporal resolution, high dynamic range(HDR), sparse and minimal motion blur. Recently, a lot of computer vision approaches have been proposed with demonstrated success. However, there is a lack of some general methods to expand the scope of the application of event-based vision. To be able to effectively bridge the gap between conventional computer vision and event-based vision, in this paper, we propose an adaptable framework for object detection in event-based vision. Our framework includes five parts: 1. Event to frame: converting event-based data to special frames for convolution processing, 2. Transfer learning: transferring the rich features learned in the above step to event-based vision, 3. E-model: enhancing the detail of the event-based data, 4. C-model: compressing and filtering event-based data, 5. Detector: to better inherit the achievement of computer vision, we adopted several conventional detection algorithms as detectors in the object detection process. Part 1 and part 2 are steps of preprocessing in our framework, parts 3-5 are the main steps of a detection algorithm for event-based data in our framework, E-model, and C-model overcomes the limitations of dense noise and sparse object in event-based data. Additionally, we collected a small event camera dataset from the real world to confirm our ideas and evaluate the AFOD framework in the object detection task. The experiments show that our proposed approach is very competitive and extensible. Specifically, our basic framework can be extended to other learning domains, to the best of our knowledge, there is no obstacle to incorporate other deep learning models in our proposed framework for event-based vision.

Index Terms—Event-based vision, adaptable framework, object detection

I. INTRODUCTION

IN recent years, deep learning methods operating on frame-based data have reached remarkable results with large-scale data training [1]–[5]. The convolutional neural networks (CNNs) have significantly improved the performance of visual classification and object detection [7]–[10], and the rich semantic and structural features could be extracted from the trained CNN models, which trained by large-scale datasets. These successful applications and datasets are based on frame-based cameras or CMOS RGB cameras. However, with the development of sensor and chip technology, more and more sensing devices have been developed to improve the ability to capture information in various environments. For example,

in self-driving car [11] [6], it integrates a large number of different types of sensors such as LIDAR, video camera, and PADRA [16]. Event camera [12]–[14], [17]–[22] is a new type of sensor adapted to the moving scene that has been proposed in recent years. In this article, we have made a rewarding research attempt on event-based vision, our method fully take into account the advantages of deep learning and the characteristics of event cameras.

Event-based vision can be thought of as a new approach of computer vision or robot vision and captures vision data through event cameras. An event camera, such as silicon retina, neuromorphic camera, or dynamic vision sensor (DVS), is a bio-inspired sensor that updates and responds to every pixel change in brightness. In addition, event cameras have lower overexposure and 120 dB dynamic range than conventional frame-based cameras [12].

Event cameras have some advantages compared to frame-based cameras. These advantages will make it play a greater role in some technical fields such as self-driving car, drone, 3D reconstructed and SLAM [6] [23]–[25]. In this part, we will give several outstanding characteristic of event cameras:

High dynamic range: Event cameras have a high dynamic range above 120 db. They can capture effective information

This work was supported by the Science and Technology Development Fund (FDCT) of Macau under Grant No. 0016/2019/A1.

Shixiong Zhang is with the International Institute of Next Generation Internet, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, China (e-mail: zsxpascal@gmail.com).

Wenmin Wang, corresponding author, is with the International Institute of Next Generation Internet, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, (e-mail: wmwang@must.edu.mo).

Honglei Li is with the International Institute of Next Generation Internet, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, China.

Fig. 1. Illustration of the Frame-based vision vs Event-based vision, Frame-based vision is updated by frame, and Event-based vision is updated by asynchronous pixel. For event-based data, it seems to be dense in time series, but it is sparse in space, which is one of the main characteristics of event-based data.

when the light changes drastically. It is due to the fact that each pixel in the event camera works independently without waiting for the global shutter.

No motion blur: The frame of a frame-based camera is likely to be captured with motion blur when it is capturing a fast-moving target or the camera shakes. But event cameras can capture very fast motions without motion blur. One of the main reasons is that every event pixel is changed and updated with microsecond resolution, which leads to a very high temporal resolution.

Ultra-low power: Each pixel of the event camera is independent and asynchronous. When the brightness changes, it will only update the local pixels instead of all the global pixels. Therefore, the event camera has very low power consumption. The average power of the event camera is 1 mW.

Low latency: The latency of a high-speed frame-based camera with good performance is 1ms. However, the latency of an event-based camera blows 0.1 ms. This allows it to be applied to fast-moving self-driving cars or drones.

Despite the remarkable characteristic of event cameras, the application of event cameras is still limited by algorithms, datasets, and tools. Each year, thousands of algorithms are proposed in computer vision, from this, it's a priority that we should develop event camera methods based on many algorithms in computer vision, especially object detection.

In this paper, we design an adaptable framework that can be used to object detection task on event cameras. Our purpose is to modify the current computer vision algorithm to adapt to the event-based vision. The main contributions of this paper are as follows:

- We proposed an AFOD framework for object detection of event-based data, which includes five parts: event to frame, transfer learning, E-model, C-model, and detector. Since the detector can be any one of object detection

models, thus other detectors are easily adaptable our AFOD framework for event-based vision, as shown in Fig.2 ;

- We adopt five models of three different types for object detection: transformer-based method, anchor-based methods (YOLOV3, YOLOV4) , and anchor-free methods (CenterNet, FCOS) as detectors in our framework respectively. And we find the transformer-based method to the event-based vision is demonstrated best performance, as shown in Fig.3, Table 1, Table 2;
- We gathered a small event-based dataset from the real world. We have verified through our own dataset that our AFOD framework can work well with the limited event-based dataset. Through our AFOD framework, we provide a novel approach for expanding the application of event-based vision without large-scale event-based datasets.
- We experiment that rich features learned by deep learning networks can be well transferred to serving event-based vision. In order to better verify our method, we designed ablation experiments in detail and experimental results show that our AFOD to successfully address the task of object detection and demonstrates strong competitiveness and expandability in event-based vision.

We have organized the rest of this paper in the following way: Section II describes the related work on object detection and event-based vision, in this part, we will review recent research papers. Section III introduces our AFOD framework and presented the details of transformer based method as detector for event-based vision. Section IV shows our experiments and ablations, and we will give discussions of our observations. The paper ends with conclusions and we also propose our future research works (Section V).

II. RELATED WORK

In this section, we will firstly review several object detection methods in computer vision. And then, we will review the latest research progress and applications of event-based vision.

A. Object detection

Object detection is an important research topic in the field of computer vision with numerous applications. So far, many important algorithms and models have been proposed [15], [26]–[28] [32]. There are various typical models of visual object detection. Most object detection algorithms are can be summarized into the following types: anchor-based methods and anchor-free methods.

Anchor-based methods: Object detection based on anchor has been for several years [36]–[39]. In another aspect, all one-stage detection algorithms and two-stage detection algorithms belong to the anchor-based detectors due to them sample a large number of preset anchors on the frame. Two-stage detection algorithms are just as their names imply that one requires two processing to complete the detection task, one-stage detection algorithms only requires one step.

Kaiming He and his research partners have made important contributions to proposed a set of two-stage models. They presented a series of R-CNN models to promote the development

of the field of object detection, including R-CNN [32], Fast R-CNN [33], Faster R-CNN [35], and Mask R-CNN for Image segmentation [34]. These models are highly correlated, and the new version shows great performance and speed improvement compared to the old ones. The main ideas of these R-CNN models are that thousands of object region candidates are selected through selective search (2k candidates per frame), and then the CNNs features are extracted from each region candidate individually for object detection. This type of model can achieve high detection accuracy but could be too slow for some real-time applications since they need additional models to provide region proposals.

For one-stage detection algorithms, this type of detection algorithm directly skips the region proposal stage, they don't need thousands of candidate regions and only predict a limited number of bounding boxes. Although the accuracy is slightly lost compared to the two-stage detectors, it can run super-fast. Here are some hot one-stage detection algorithms such as YOLO family [28]–[31], SSD [45]–[47]. No region proposal, YOLO and SSD only run a single convolutional network on the frame.

Anchor free methods: Nowadays, anchor-free methods [40]–[44] attracted much attention because they achieve excellent performance in the object detection task. In contrast to anchor-based methods, anchor-free methods don't need to set up a large number of hand-designed anchors. They optimize the deep network of object detection using the proposal of FPN and Focal Loss. These anchor-free methods can eliminate anchor-related hyperparameters, so that, they have more potential in generalization ability. In the following, we will describe two state-of-the-art anchor-free methods: CenterNet [48] and FCOS [49], both of these algorithms achieve much better performance than anchor-based methods.

CenterNet is exploring the neural network to predict the objects and their location without anchors. They model an object as a key point, and the key point is the center point of the bounding box. Their methods find the object attributes such as location, size, direction, and 3D position by evaluating the keypoints. Additionally, CenterNet can be used not only for object detection but also for other tasks, such as body detection or 3D bounding box detection, etc.

FCOS is a one-stage detector that absolutely avoids the very difficult computation related to anchor boxes. These anchors usually contain a large number of hyperparameters. FCOS uses regression to measure the center-ness of positive sample location in the bounding box. Most detectors trained one multi-class classifier for object classification, but FCOS trained C binary classifiers, where C denotes the number of object categories. These steps effectively improve the performance of the anchor-free detector.

B. Event-based vision

Event cameras work fundamentally differently from frame-based cameras. Although event-based vision hasn't been presented for a long time, the first commercial device: Dynamic Vision Sensor (DVS) was produced by iniVation in 2008. With the continuous efforts of researchers, event cameras have been applied in many fields, such as feature extraction

and tracking, optical flow estimation, depth estimation, pose estimation and SLAM, VIO, image reconstruction, motion segmentation, recognition, neuromorphic control, etc. Research in this field is very active, and new algorithms are constantly being produced [59]. The latest Celex-V camera [51] is the world's first-megapixel event camera, this camera is used in our experiments. Due to the special characteristics of event cameras, most traditional computer vision algorithms cannot be directly applied to event cameras, so new frameworks are demanded for event cameras.

In many studies on event-based vision, two types of methods are commonly extended by researchers. One of the methods is to use the correct parameters to convert the event-based data into frame data and then process it. For instance, Gallego et al [52] try to use the correct parameters to convert the event to the reference frame, and to address several computer vision problems with event cameras: motion, depth, and optical flow estimation. Similar methods are also proposed in paper [53] and paper [54]. Another method is to use SNN to process event camera data [55]–[57]. SNN is standing for spiking neural networks. Similar to the principle of an event camera, SNN is an artificial neural network inspired by the brain. SNNs show their potential as new types of artificial networks, especially for processing event-based data from neuromorphic sensors. So far, the research and application of event-based vision is a constant increase, such as image deblurring [58] or star tracking [59], which will be displayed as an event camera becomes widely available.

III. METHODOLOGY

In this section, we will introduce our methodology, we first present how to convert event data into frame data. In order to learn rich event features, we presented a special feature transfer learning approach. And then, our AFOD framework will be described in detail. We propose our AFOD framework is based on the characteristics of event data and our framework can adopt various algorithms as detectors to address the task of object detection in event vision. In this section, we developed transformers model to summarize the global information from CNNs features. In order to highlight the structural features of the object from the sparse event-based data, we have designed an E-model to enhance the event-based data. Since the event data is enhanced by E-model, the background noise has also been enhanced in the event data, so we adopted a compression and sparse method to process event-based data. The amount of background noise is far more than the amount of object in the event data, a lot of background noise is compressed out in the process of compression. Although the amount of data is reduced, the data of the object structure is well preserved. Our framework can be divided into 5 main parts as shown Fig.2, since each step of our framework is relatively independent, extending some conventional algorithms to our AFOD framework should be uncomplicated in event-based vision, as shown Fig.3 .

A. Event to frame

In computer vision, most methods are successfully demonstrated by convolution processing frame data. As a special

Fig. 2. The main steps of our framework, the detector can be adopted with a variety of conventional detection algorithms. Therefore there is always a lag behind the most recently proposed detectors for them to be incorporated in our proposed framework.

Fig. 3. Overview of the proposed method. We adopted a transformer model as our detector. This diagram shows the main steps of our AFOD. The backbone was trained by a large scale RGB dataset and they can learn rich structure features from this RGB dataset. So, they will be able to effectively extract event object features when they are retrained by the Small DVS dataset, where the Small DVS dataset denotes our event-based dataset.

camera, the event camera has the ability to capture event data with image characteristics. One of the most effective ways to exploit event data is to convert event data into frame data [60]–[66], thus the event-based frame can retain spatial information. Different from the synchronous frame captured by CMOS cameras one frame after another, the event camera captures asynchronous time series data [70]–[73]. We denoted a sequence of event data as $E = \{(x_i, y_i, p_i, t_i) | i \in n\}$, where x and y represent the spatial coordinates of the event. p denotes the event's intensity, and the meaning of t is a timestamp. Because the event camera is updated very quickly, we can gather all event pixel as a frame in a fixed time period ∇t , we generate event frames F of size $H \times W$ as:

$$F_1, F_2, F_3, \dots, F_n \quad (1)$$

We consider two functions: $f_t(\cdot)$ as temporal function, $f_s(\cdot)$ as spatial function, the aggregation of event-based data in a fixed time period ∇t is: $f_s(f_t(\cdot))$. Each frame is defined as:

$$F = f_s(f_t(E)) \quad (2)$$

we normalize the event's intensity p_w as the pixel value of F : $p_w \in \{0 : 255\}$.

B. Transfer Learning

According to the experience of object detection on RGB image datasets, an effective way to improve the performance is to train a deep convolutional neural network model on large-scale datasets. However, so far, there are no real large-scale event camera datasets. In the previous section, we proposed a method to convert event data into special frame data, as shown in Fig.4. The clear structure of the object can be observed,

Fig. 4. Illustration of the visual image that converts the event data into frame data. These images show the object in the traffic scenarios, these data will support event cameras to expand their applications in self-driving cars and smart cities.

although the event data is sparse. So that, it is possible that rich object structure features can be learned from large-scale RGB image datasets, and these features can be transferred to event-based vision through transfer learning [67]–[69].

In this article, we provide a transfer learning approach to bridge the gap between image feature extraction and event feature extraction. We developed CNNs as our backbone networks to extract object features. We formalize our method

as: Given a learning task T_r with source domain D_r , and we have a target domain D_e and a learning task T_e . The purpose of this step is to improve the learning of the prediction function $f_e(\cdot)$ of the target domain. We consider RGB frame-based learning as the source domain, and event frame-based learning as the target domain, the structure feature can be learned by optimizing a supervised learning regularization error function, given as follows:

$$\arg \min_{M, V} \sum_{t \in (D_r, D_e)} \sum_{i=1}^n L(y_{ti}, \langle a_{ti}, \mathbf{V}^\top X_{ti} \rangle) + \lambda \|\mathbf{M}\|_{2,1}^2, \quad (3)$$

Where D_r and D_e denote the supervised learning tasks in the source domain and target domain, respectively. $\mathbf{V}$ means that an orthogonal matrix can map the original high-dimensional data to low-dimensional information, $\mathbf{M}$ is parameter matrix, λ denote the regularization parameter, and a_t denote the learned vector. Especially, the main purpose is to learn an optimal parameter p^* by minimizing an expected risk $\mathbb{E}$.

$$p^* = \arg \min_{p \in \Theta} \mathbb{E}(\phi, \omega) \in d[l(\phi, \omega, p)], \quad (4)$$

where l is a loss function, ϕ is a particular sample of data, the label is defined as ω , d is a probability distribution.

C. E-model

In this part, we will introduce the our enhancement model (E-model) in our framework. The main function of the E-model is to enhance the event-based data. One of the main characteristics of event-based data is sparseness, as shown in Fig.4. Image enhancement is widely needed in some areas with sparse data, such as medical image processing, astrophotography, processing of remote sensing image data. Although deep learning networks based image enhancement methods are hot research topic, it is not suitable for event-based data. Through our analysis and experiment, we adopted a simple and effective traditional model to enhance the event-based data. Our enhancement model is based on Gamma correction [80]–[82] which is an efficient and simple image contrast enhancement tool.

$$U(p_w) = \text{round} \left[p_{w_{max}} \left(\frac{p_{w_{max}}}{p_w} \right)^\gamma \right] \quad (5)$$

where $U(p_w)$ is the output value, $\text{round}[\cdot]$ is the rounding operation. p_w represents the event pixel value, γ is the gamma value. The value of gamma γ is considered to be less than 1, therefore, most low-value pixels in the event-based frame will be stretched. As a result, the pixel dynamic range of the event-based frame will be expanded, as shown in Fig.5.

D. C-model

In this part, we proposed our compression model (C-model) to compress and filter event-based data. The C-model is very based on the Gaussian process. Since the event-based data contains a lot of background noise. In the previous subsection, we presented an E-model to enhance the event-based data, however, the E-model not only enhanced the structure and

Fig. 5. Illustration of the intensity histogram, the picture below shows more details than the picture above

details of the object but also increased a large amount of background noise.

On the grounds that the sparse characteristic of event data, a large amount of background noise will cause data redundancy. In order to reduce noisy data and data redundancy, we propose a sparse filter based on a Gaussian template. It is closely related to the Gaussian distribution, the problem can be formulated as:

$$\mathbf{v} = R\mathbf{f}. \quad (6)$$

where R is a Gaussian matrix, $R \in \mathbf{R}^{n \times m}$, $\mathbf{f}$ is a high-noise event-based frame space, and $\mathbf{v}$ is a lower-noise event-based frame space, a two-dimensional Gaussian function is as follows :

$$r(i, j) = e^{-\frac{i^2 + j^2}{2\sigma^2}} \quad (7)$$

We assume that the size of Gaussian sparse filter matrix is $(2k+1) \times (2k+1)$

$$R(i, j) = \frac{1}{2\pi\sigma^2} e^{-((i-k-1)^2 + (j-k-1)^2)/2\sigma^2} \quad (8)$$

We use Gaussian templates as sparse filters, it has the ability to play an important role in event-based data. High-frequency noise can be effectively sparse, however, the structure of the object can be well preserved. It is important that we can decrease data redundancy and noise without performance reduction, it will show clearly and verification in ablation study.

E. Transformer

The transformer has been proven to be able to extract global information well due to its unique structure, our experiments show that its special network is very beneficial for dealing with sparse event-based data. We adopt a transformer as a detector to achieve the best result in our experiments. We will describe the transformer in this part.

Fig. 6. Illustration of the Comparison Chart, The number of C1, C2, C3 is nearly half of the number of F1, F2, F3, respectively

The transformer is a novel NLP model proposed by the Google team in 2017, and now the more popular BERT [77]–[79] is also based on transformer. The transformer model uses the multi-head self-attention mechanism to replace several sequential structures such as RNN, LSTM, and GRU so that the model can be trained in parallel and have benefits for extracting global information. The transformer is a Sequence-to-Sequence model which includes several query elements and key elements. The inspiration originates from vision transformer [74]–[76], we adopted transformer in event-based object detection. The task of object detection is to predict the coordinates of a series of bounding boxes and labels. Most modern detectors overcome this task indirectly by defining some proposals, anchors, or windows and constructing the problem into a classification and regression problem. However, transformer can directly predict the final detections (class and bounding boxes) without a set of hand-designed components such as region proposal, anchor, Non-Maximum Suppression (NMS), and so on. The main steps of the transformer are shown in Fig. 7. We consider three inputs: feature map, positional encoding and object queries. The feature map f_d is extracted by CNNs backbone, we consider it as $f_d = CNNs(f)$. Since the transformer model is parallel and independent, the position encoding uses sine and cosine functions to provide the necessary position information. Object queries provide N learnable embeddings to produce N bounding boxes and classes. The multi-head self-attention mechanism can be formulated as:

$$MultiAtt(f_q, f_k) = \sum_{n=1}^N W_n \left(\sum_{k \in \Omega_k} A_{nqk} \cdot W_n' f_k \right) \quad (9)$$

Where f_q is query element, and f_k are a set of key element, A_{nqk} are attention weights, W_n and W_n' are both learnable weight. The different attention heads allow model to attend to contents from various representation subspace and different

Fig. 7. Illustration of our transformer-based model, input CNNs features maps and positional encoding to transformer encode, during training, a transformer decode directly predicts a set of prediction results, object queries is fixed number of learned positional embedding.

positions, more details in [74], [75], [84]–[87].

IV. EXPERIMENTS

There are not many available algorithms and datasets for event-based object detection. In this section, we provide a dataset to evaluate our proposed framework. So far, event cameras have great potential to expand their applications in the fields of autonomous driving and smart cities, so we collected event-based data from real traffic scenes. The Big RGB dataset is the COCO-train2017 dataset. In an ablation study, we first evaluate the main peripherals of our approach, and then, we extend recent state-of-the-art object detection algorithms to our proposed frameworks and compared them with each other. Setup: The proposed AFOD on the Intel Xeon(R) Silver 4114 CPU@2.20GHz x 40 and GeForce RTX2080Ti GPU. All referenced algorithms are implemented with recommended parameters and use ResNet101 as the backbone which has been trained with the same large-scale datasets.

A. Event-based dataset

DATASETS: Since the event camera has not yet been mass-produced and applied on a large scale, real event camera data is always scarce. Most event camera-based algorithms are trained and tested based on artificially simulated data. Our small event-based dataset collected from real events is used to test and verify our framework. We gathered data and reconstructed frames on these data, and we set ∇t as 30ms. The mumble of an event-based dataset is 2.6k, including 2k training data, and 600 test data. The main object to be detected is the moving vehicle on the road. We consider ResNet101 as the backbone, the experiments show that our proposed approach is very competitive in real event scenes.

B. Ablation Study

In our experiments, we use conventional object evaluation criteria: AP (Average precision) to evaluate our methods. Average precision calculates the value of average precision for recall values over 0 to 1. In this part, we evaluate the importance of the key steps of our approach. As shown in Table 1, our framework adopt transformer method achieves 69.9 $AP_{0.5:0.95}$, E-model has been shown to effectively improve performance under the same conditions, $AP_{0.5:0.95}$ increases by 6.9%. Fig.6 provides a diagram of intensity histogram, the first graph has a smooth curve that donates the source event-based frame, and the second graph shows that the event-based frame shows more details for the object after it is enhanced by E-model. Similarly, we also try to apply CNNs based enhancement [83] to strengthen event-based data, it did not work well because it was not trained with enough event-based data, as shown in the fourth row of table 1. We use a histogram to describe the change in the number of pixels after C-model processing, F1, F2, F3 show three example event-based frames, C1, C2, C3 show the results of applying the C-model on F1, F2, F3, respectively, as Fig.6 shows. It should be noticed that the number of event-based data is reduced by nearly half, the performance of our proposed method has not crumbled, but performance has improved in some indicators. It proves that our method can effectively reduce the proportion of noisy data and retain structure features information of the object, as Fig.6 and Table 1 show.

In the second set of experiments, we extend some popular object detection algorithms to our frameworks for comparison fair. Except for transformer model, we adopted anchor-based detections (YOLOV3, YOLOV4) and anchor free detections (CenterNet, FCOS) in our framework. We develop all our main steps for these algorithms, as shown in Table 2. All of these algorithms use the same pretrained backbone and test on the same dataset. Here, we proposed both Big RGB dataset and Small DVS dataset for YOLOV4(1), only Big RGB dataset for YOLOV4(2), and only Small DVS dataset for YOLOV4(3). It should be noted that the performance reaches 0 in YOLO4(2), showing that the step of retraining in small event-based data is necessary, and the features of event-based data are similar but different from features of RGB-based data. As we can see in YOLO4(1) and YOLOV4(3), we show the advantage of using the transfer learning in Sec III-B, when learning the rich features in large-scale RGB images dataset, $AP_{0.5:0.95}$ increases by 14.3%, and $AP_{0.75}$ increase by 9.9%. This proves that the rich object structure features can be learned from large-scale RGB image datasets, and these features can be transferred to event-based vision through transfer learning. It should be noted that the transformer-based detector achieves better performance due to its powerful global information extraction capabilities in small event-based data. The advantage of the transformer model is easier to observe in sparse event-based data.

V. CONCLUSION

In this article, we first present an adaptable framework (AFOD) that is able to bridge the gap between conventional

computer vision and event-based vision. Our AFOD for object detection in event-based vision which includes: event to frame, transfer learning, E-model, C-model, and detector. And then, a small event-based dataset has been gathered from the real world by using event-camera Celex-V. We applied this event-based dataset to train and test our proposed framework. Especially, event to frame means that we convert event data into frame data for convolution processing. Limited by small training event-based datasets, we proposed transfer learning to learn rich features from large-scale RGB datasets. Additionally, we proposed an E-model and C-model to improve the performance based on the characteristic analysis of event-based data. The detector means that different types of conventional detection algorithms are easy to extend in event-based vision through our AFOD. Experimental results show that the effectiveness of our proposed AFOD for event-based object detection has been successfully demonstrated. Through our research, we have fully understood the characteristics of the event-based vision. Although its data is very sparse, it contains sparse recursive representations of the object, and the transformer model shows great potential in handling event-based data. We hope that our research will be constructive for other researchers in event-based vision.

In future work, we will more fully verify our method on various event-based datasets and extend our framework to another research topics such as event-based object tracking, event-based object segmentation, and so on. Based on the unique characteristics of event-based vision, we look forward to expanding more applications in event-based vision, allowing it to play an important role in real systems.

REFERENCES

- [1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009.
- [2] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115.3 (2015): 211-252.
- [3] Noh, Hyeonwoo, et al. "Large-scale image retrieval with attentive deep local features." Proceedings of the IEEE international conference on computer vision. 2017.
- [4] Radenović, Filip, et al. "Revisiting oxford and paris: Large-scale image retrieval benchmarking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
- [5] Niu, Yulei, et al. "Multi-modal multi-scale deep learning for large-scale image annotation." IEEE Transactions on Image Processing 28.4 (2018): 1720-1731.
- [6] Maqueda, Ana I., et al. "Event-based vision meets deep learning on steering prediction for self-driving cars." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
- [7] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012): 1097-1105.
- [8] Tan, Mingxing, Ruoming Pang, and Quoc V. Le. "Efficientdet: Scalable and efficient object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
- [9] Yin, Tianwei, Xingyi Zhou, and Philipp Krahenbuhl. "Center-based 3d object detection and tracking." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- [10] Zhao, Zhong-Qiu, et al. "Object detection with deep learning: A review." IEEE transactions on neural networks and learning systems 30.11 (2019): 3212-3232.
- [11] Badue, Claudine, et al. "Self-driving cars: A survey." Expert Systems with Applications (2020): 113816.

TABLE I

ABLATION STUDY ON OUR TRAFFIC EVENT DATASET. THERE ARE THREE MAIN COMPONENTS OF OUR METHOD: E-MODEL, C-MODEL, AND TRANSFORMER MODEL. THE BEST RESULTS ARE HIGHLIGHTED BY BOLD FACED

Methods	E-model	C-model	$AP_{0.5:0.95}$	$AP_{0.5}$	$AP_{0.75}$	AP_S	AP_M	AP_L
AFOD+Transformer			63.0	95.4	74.5	45.0	64.5	66.1
		✓	60.3	92.0	69.8	34.2	64.1	61.2
	✓		69.7	96.5	85.3	53.2	71.2	71.5
	(CNNs)	✓	63.0	95.3	77.3	44.6	64.5	65.7
	✓	✓	69.9	96.3	84.1	49.1	71.8	72.1

TABLE II

EXTEND STATE-OF-ART OBJECT DETECTION ALGORITHMS TO OUR FRAMEWORK AND COMPARED THEM WITH EACH OTHER. THE BEST RESULTS ARE HIGHLIGHTED BY BOLD FACED

Detectors	Anchor-free	Big RGB dataset	Small DVS dataset	ResNet-101	$AP_{0.75}$	$AP_{0.5:0.95}$
AFOD+YOLOV3 [30]		✓	✓	✓	58.9	55.7
AFOD+YOLOV4(1) [31]		✓	✓	✓	79.6	64.1
AFOD+YOLOV4(2) [31]		✓		✓	0.00	0.00
AFOD+YOLOV4(3) [31]			✓	✓	66.7	49.8
AFOD+CenterNet [48]	✓	✓	✓	✓	79.2	65.8
AFOD+FCOS [49]	✓	✓	✓	✓	83.8	67.4
AFOD+Transformer [75]	✓	✓	✓	✓	84.1	69.9

- [12] G. Gallego et al., "Event-based Vision: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2020.3008413
- [13] Rusci, Manuele, et al. "An event-driven ultra-low-power smart visual sensor." IEEE Sensors Journal 16.13 (2016): 5344-5353.
- [14] Chen, Guang, et al. "A novel visible light positioning system with event-based neuromorphic vision sensor." IEEE Sensors Journal 20.17 (2020): 10211-10219.
- [15] Zhang, Shuai, et al. "New object detection, tracking, and recognition approaches for video surveillance over camera network." IEEE sensors journal 15.5 (2014): 2679-2691.
- [16] Zhao, Xiangmo, et al. "Fusion of 3D LIDAR and camera data for object detection in autonomous vehicle applications." IEEE Sensors Journal 20.9 (2020): 4901-4913.
- [17] Benosman, Ryad, et al. "Event-based visual flow." IEEE transactions on neural networks and learning systems 25.2 (2013): 407-417.
- [18] Lichtsteiner, Patrick, Christoph Posch, and Tobi Delbruck. "A 128×128 120 dB $15 \mu s$ latency asynchronous temporal contrast vision sensor." IEEE journal of solid-state circuits 43.2 (2008): 566-576.
- [19] Brandli, Christian, et al. "A 240×180 130 db $3 \mu s$ latency global shutter spatiotemporal vision sensor." IEEE Journal of Solid-State Circuits 49.10 (2014): 2333-2341.
- [20] Muglikar, Manasi, et al. "How to Calibrate Your Event Camera." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- [21] Chen, Guang, et al. "Event-based neuromorphic vision for autonomous driving: a paradigm shift for bio-inspired visual sensing and perception." IEEE Signal Processing Magazine 37.4 (2020): 34-49.
- [22] Chen, Guang, et al. "EDDD: Event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor." IEEE Sensors Journal 20.11 (2020): 6170-6181.
- [23] Kim, Hanme, Stefan Leutenegger, and Andrew J. Davison. "Real-time 3D reconstruction and 6-DoF tracking with an event camera." European Conference on Computer Vision. Springer, Cham, 2016.
- [24] Carneiro, João, et al. "Event-based 3D reconstruction from neuromorphic retinas." Neural Networks 45 (2013): 27-38.
- [25] Vidal, Antoni Rosinol, et al. "Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios." IEEE Robotics and Automation Letters 3.2 (2018): 994-1001.
- [26] Viola, Paul, and Michael Jones. "Robust real-time object detection." International journal of computer vision 4.34-47 (2001): 4.
- [27] Kang, Kai, et al. "T-cnn: Tubelets with convolutional neural networks for object detection from videos." IEEE Transactions on Circuits and Systems for Video Technology 28.10 (2017): 2896-2907.
- [28] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [29] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- [30] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).
- [31] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).
- [32] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
- [33] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.
- [34] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.
- [35] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031
- [36] Kong, Tao, et al. "Foveabox: Beyond anchor-based object detection." IEEE Transactions on Image Processing 29 (2020): 7389-7398.
- [37] Zhang, Xiaosong, et al. "Learning to match anchors for visual object detection." IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).
- [38] Zhong, Yuanyi, et al. "Anchor box optimization for object detection." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.
- [39] Li, Shuai, et al. "Dynamic anchor feature selection for single-shot object detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
- [40] Zhu, Chenchen, Yihui He, and Marios Savvides. "Feature selective anchor-free module for single-shot object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
- [41] Lee, Youngwan, and Jongyoul Park. "Centermask: Real-time anchor-

- free instance segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
- [42] Yan, Yichao, et al. "Anchor-Free Person Search." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- [43] Li, Jiachen, et al. "Pseudo-IoU: Improving Label Assignment in Anchor-Free Object Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- [44] Chen, Qi, et al. "Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots." European Conference on Computer Vision. Springer, Cham, 2020.
- [45] Liu, Wei, et al. "Ssd: Single shot multibox detector." European conference on computer vision. Springer, Cham, 2016.
- [46] Fu, Cheng-Yang, et al. "Dssd: Deconvolutional single shot detector." arXiv preprint arXiv:1701.06659 (2017).
- [47] Zhang, Shifeng, et al. "Single-shot refinement neural network for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
- [48] Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." arXiv preprint arXiv:1904.07850 (2019).
- [49] Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
- [50] Chin, Tat-Jun, et al. "Star tracking using an event camera." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.
- [51] Chen, Shoushun, and Menghan Guo. "Live demonstration: CeleX-V: a 1M pixel multi-mode event-based sensor." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2019.
- [52] Gallego, Guillermo, Henri Rebecq, and Davide Scaramuzza. "A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
- [53] Stoffregen, Timo, and Lindsay Kleeman. "Event cameras, contrast maximization and reward functions: an analysis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
- [54] Gallego, Guillermo, Mathias Gehrig, and Davide Scaramuzza. "Focus is all you need: Loss functions for event-based vision." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
- [55] Gehrig, Mathias, et al. "Event-based angular velocity regression with spiking networks." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.
- [56] Lichtsteiner, Patrick, and Tobi Delbruck. "A 64x64 AER logarithmic temporal derivative silicon retina." Research in Microelectronics and Electronics, 2005 PhD. Vol. 2. IEEE, 2005.
- [57] Chan, Vincent, Shih-Chii Liu, and Andr van Schaik. "AER EAR: A matched silicon cochlea pair with address event representation interface." IEEE Transactions on Circuits and Systems I: Regular Papers 54.1 (2007): 48-59.
- [58] Pan, Liyuan, et al. "Bringing a blurry frame alive at high frame-rate with an event camera." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
- [59] Chin, Tat-Jun, et al. "Star tracking using an event camera." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.
- [60] Scheerlinck, Cedric, et al. "Fast image reconstruction with an event camera." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.
- [61] Duwek, Hadar Cohen, Albert Shalumov, and Elishai Ezra Tsur. "Image Reconstruction From Neuromorphic Event Cameras Using Laplacian-Prediction and Poisson Integration With Spiking and Artificial Neural Networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- [62] Wang, Lin, Tae-Kyun Kim, and Kuk-Jin Yoon. "Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
- [63] Mitrokhin, Anton, et al. "Event-based moving object detection and tracking." 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.
- [64] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in Int. Conf. Comput. Vis. (ICCV), 2019.
- [65] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "EV-FlowNet: Self-supervised optical flow estimation for event-based cameras," in Robotics: Science and Systems (RSS), 2018.
- [66] S. Tulyakov, F. Fleuret, M. Kiefel, P. Gehler, and M. Hirsch, "Learning an event sequence embedding for dense event-based deep stereo," in Int. Conf. Comput. Vis. (ICCV), Oct. 2019.
- [67] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.
- [68] Evgeniou, An, and Massimiliano Pontil. "Multi-task feature learning." Advances in neural information processing systems 19 (2007): 41.
- [69] Wu, Yue, and Qiang Ji. "Constrained deep transfer feature learning and its applications." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [70] Duo, Jingyun, and Long Zhao. "An Asynchronous Real-Time Corner Extraction and Tracking Algorithm for Event Camera." Sensors 21.4 (2021): 1475.
- [71] Gehrig, Daniel, et al. "End-to-end learning of representations for asynchronous event-based data." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
- [72] Bardow, Patrick, Andrew J. Davison, and Stefan Leutenegger. "Simultaneous optical flow and intensity estimation from an event camera." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
- [73] Rodríguez-Gómez, Juan Pablo, et al. "Asynchronous event-based clustering and tracking for intrusion monitoring in uas." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.
- [74] Zhu, Xizhou, et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection." ICLR 2021.
- [75] Carion, Nicolas, et al. "End-to-end object detection with transformers." European Conference on Computer Vision. Springer, Cham, 2020.
- [76] Zhou, Daquan, et al. "Deepvit: Towards deeper vision transformer." arXiv preprint arXiv:2103.11886 (2021).
- [77] Devlin, J. et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT (2019).
- [78] Polignano, Marco, et al. "Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets." 6th Italian Conference on Computational Linguistics, CLiC-it 2019. Vol. 2481. CEUR, 2019.
- [79] Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. "A primer in bertology: What we know about how bert works." Transactions of the Association for Computational Linguistics 8 (2020): 842-866.
- [80] Farid, Hany. "Blind inverse gamma correction." IEEE transactions on image processing 10.10 (2001): 1428-1433.
- [81] Huang, Shih-Chia, Fan-Chieh Cheng, and Yi-Sheng Chiu. "Efficient contrast enhancement using adaptive gamma correction with weighting distribution." IEEE transactions on image processing 22.3 (2012): 1032-1041.
- [82] Rahman, Shanto, et al. "An adaptive gamma correction for image enhancement." EURASIP Journal on Image and Video Processing 2016.1 (2016): 1-13.
- [83] Wei, Chen, W. Wang, W. Yang and Jiaying Liu. "Deep Retinex Decomposition for Low-Light Enhancement." BMVC (2018).
- [84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)
- [85] Parmar, Niki, et al. "Image transformer." International Conference on Machine Learning. PMLR, 2018.
- [86] Floridi, Luciano, and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." Minds and Machines 30.4 (2020): 681-694.
- [87] Srinivas, Aravind, et al. "Bottleneck transformers for visual recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.