Retinomorphic Object Detection in Asynchronous Visual Streams

Jianing Li[1*], Xiao Wang[3*], Lin Zhu[1], Jia Li[2,3], Tiejun Huang[1], Yonghong Tian[1,3†]

[1]National Engineering Laboratory for Video Technology, School of Computer Science, Peking University, Beijing, China
[2]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University, Beijing, China
[3]Peng Cheng Laboratory, Shenzhen, China
{lijianing, linzhu, tjhuang, yhtian}@pku.edu.cn, wangx03@pcl.ac.cn, jiali@buaa.edu.cn

Abstract

Due to high-speed motion blur and challenging illumination, conventional frame-based cameras have encountered an important challenge in object detection tasks. Neuromorphic cameras that output asynchronous visual streams instead of intensity frames, by taking the advantage of high temporal resolution and high dynamic range, have brought a new perspective to address the challenge. In this paper, we propose a novel problem setting, ***retinomorphic object detection***, which is the first trial that integrates foveal-like and peripheral-like visual streams. Technically, we first build a large-scale multimodal neuromorphic object detection dataset (i.e., PKU-Vidar-DVS) over 215.5k spatio-temporal synchronized labels. Then, we design temporal aggregation representations to preserve the spatio-temporal information from asynchronous visual streams. Finally, we present a novel bio-inspired unifying framework to fuse two sensing modalities via a dynamic interaction mechanism. Our experimental evaluation shows that our approach has significant improvements over the state-of-the-art methods with the single-modality, especially in high-speed motion and low-light scenarios. We hope that our work will attract further research into this newly identified, yet crucial research direction. Our dataset can be available at https://www.pkuml.org/resources/pku-vidar-dvs.html.

Introduction

Conventional frame-based cameras have presented some limitations for object detection in challenging scenes (e.g., motion blur, over-exposure, and low-light), resulting in a sharp drop in performance using unusable frames (Sayed and Brostow 2021). A key question still remains: *How does the human visual system perform information flows and capture an object in extreme challenging scenarios?*

Studies (Sinha et al. 2017; Roy, Jaiswal, and Panda 2019) have revealed that the visual system of primates transforms information flows through discrete action potentials or "***spikes***". Recently, two types of neuromorphic vision sensors (i.e., event-based cameras (Gallego et al. 2020) and time-based cameras (Chen et al. 2011)) have captured the interest of the computer vision community owning to the advantages over conventional frame-based cameras.

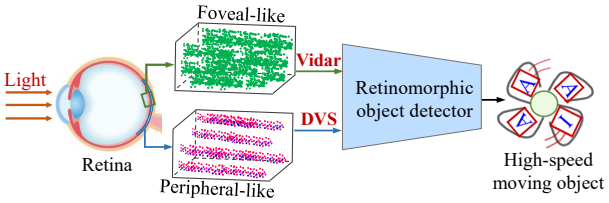


Figure 1: Retinomorphic object detector, integrating foveal-like and peripheral-like asynchronous visual streams from Vidar and DVS, performs a complementary fusion for object detection in challenging scenarios.

Event-based cameras (e.g., DVS (Lichtsteiner, Posch, and Delbruck 2008) and DAVIS (Brandli et al. 2014)) in the former type, mimicking the periphery of the retina, work in a different way over frame-based cameras: each pixel responds to intensity changes with asynchronous events. Due to natural motion sensors with a high dynamic range (HDR) and high temporal resolution, event-based cameras have been applied in object detection. Although some tasks (Perot et al. 2020; Ryan et al. 2021) achieve a satisfactory performance via only processing DVS events (i.e., brightness changes), which is hard to provide static texture (i.e., absolute brightness) and achieve high-precision performance. Another joint frameworks (Jiang et al. 2019; Hu et al. 2020; Cao et al. 2021; Liu et al. 2021; Wang et al. 2021b) attempt to combine DVS events and frames. However, a major bottleneck prohibits these joint detectors since conventional frames with 25 Hz may suffer from high-speed motion blur.

Time-based cameras (e.g., Vidar (Dong, Huang, and Tian 2017)) in the latter type, taking the foveal-like sampling model, generate a spike when the accumulation of photons for a pixel reaches a threshold. This frame-free imaging paradigm brings the ability to reconstruct visual textures using spike frequency or inter-spike interval (Zhu et al. 2020b). It has a high temporal sampling frequency of 20,000 Hz and is suitable to deal with high-speed vision tasks. Nevertheless, a slight drawback of Vidar has a lower dynamic range (70 dB versus 120 dB) over DVS, thus it is difficult to capture an object in challenging illumination scenarios.

According to the sensing preliminary from the human visual system (Stewart, Valsecchi, and Schütz 2020), processing in peripheral and foveal vision is not independent, but is more directly connected than previously thought. In other

*First two author contributed equally.
†Corresponding author: Yonghong Tian.

words, the retina is combined with the fovea and the periphery to sense real-world scenes. It motives us to ask: *Can we make complementary use of foveal-like and periphery-like visual streams for object detection as the retina does?*

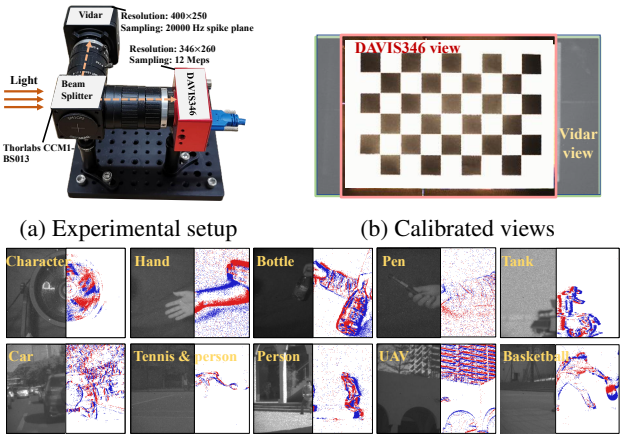
To tackle this question, we propose a novel retinomorphic object detector, which integrates foveal-like and peripheral-like asynchronous visual streams to detect objects in challenging scenarios, as shown in Fig. 1. In fact, the goal of this work is not to develop a state-of-the-art frame-based object detector with single-modality. On the contrary, we aim at overcoming ***the following challenges****: (i) How do we build a large-scale multimodal neuromorphic dataset that benefits from object detection involving high-speed and low-light? (ii) How do we leverage rich spatio-temporal cues from continuous visual streams for object detection? (iii) How does the unifying mechanism for two asynchronous streams benefit object detection?* Specifically, we first build a prototype hybrid camera system and present a large-scale multimodal neuromorphic object dataset (i.e., PKU-Vidar-DVS), which provides manual annotations at a frequency of 50 Hz for 9 classes, yielding more than 215.5k spatio-temporal synchronized labels. Then, temporal aggregation representations are proposed to preserve the spatio-temporal information from asynchronous visual streams. Motivated by the interaction between foveal and peripheral signals, a novel bio-inspired unifying object detection framework is presented to fuse two streams via dynamic interactions between sub-networks.

In summary, the main contributions of this work are:

- We introduce a novel problem setting, ***retinomorphic object detection***, which first integrates foveal-like and peripheral-like visual streams to address major object detection challenges (e.g., motion blur and low-light).
- We propose *temporal aggregation representations* using the attention mechanism, which preserves the spatio-temporal information from asynchronous visual streams.
- We present a *dynamic interaction fusion* framework via dynamically exchanging the channels for object detection, which outperforms state-of-the-art fusion methods.
- We build a *large-scale multimodal neuromorphic object detection dataset* using our hybrid camera system. We believe that this standardized dataset will open up an opportunity for the research of this challenging problem.

Retinomorphic Sensing Preliminaries

Human Retina Sensing. The retina lines the back of the eyeball that senses light and sends visual information to the brain. Central fovea with the highest resolution plays a crucial role in scrutinizing fine-detailed objects, and periphery vision is very important to identify well-known shapes and movements (Strasburger, Rentschler, and Jüttner 2011; Katzakis et al. 2019). Studies reveal that foveal and peripheral processing are closely connected and interacted to perceive the real-world scenes (Stewart, Valsecchi, and Schütz 2020). By integrating foveal and periphery vision, the retina achieves robust perception in challenging scenarios. Biological retinas have many desirable attributes which are lacking in conventional frame-based cameras but inspire the emerging neuromorphic cameras (e.g., DVS and Vidar).



(a) Experimental setup (b) Calibrated views

(c) Representative examples

Figure 2: A hybrid camera system integrates Vidar and DAVIS346. (a) A beam splitter is placed in front of two cameras with 50% splitting. (b) Spatial-temporal calibration using a standard checkerboard. (c) Examples of our PKU-Vidar-DVS dataset, the left in each image is the reconstructed image from Vidar, the right is the event image by mapping asynchronous events into the image-like plane.

Foveal-Like Sensing. Vidar, mimicking the visual sampling mechanism of the fovea, encodes the information about pixel illumination in the spike frequency or inter-spike interval (Zhu et al. 2020b; Xu et al. 2020b). One-bit spike is fired when the accumulator of the light intensity $I(x, y, t)$ for each independent pixel $[x_n, y_n]$ at the timestamp t_n reaches the threshold θ_1. Intuitively, the brighter the light, the higher frequency the spike firing, and it can be depicted as:

$$\int_0^{t_n} I(x_n, y_n, t)dt = \theta_1, \tag{1}$$

where the small integrating window dt (i.e., Δt) may result in an ultra high sampling frequency. This novel frame-free imaging paradigm enables Vidar to capture visual textures (i.e., *static information*) for ultra high-speed vision tasks.

Periphery-Like Sensing. DVS, modeling a simplified structure of the periphery, responds to light changes with asynchronous events (Posch et al. 2014; Wu et al. 2019; Zheng et al. 2021). An event e_n=(x_n, y_n, t_n, p_n), using the **a**ddress **e**vent **r**epresentation (AER) protocol (Boahen 2000), is generated once the logarithmic light change for a pixel $[x_n, y_n]$ exceeds the threshold θ_2, and it can be formulated as:

$$\ln I(x_n, y_n, t_n) - \ln I(x_n, y_n, t_n - \Delta t_n) = p_n\theta_2, \tag{2}$$

where the polarity $p_n \in \{1, -1\}$ denotes ON or OFF event, which represents the increasing or decreasing change in the brightness, and Δt_n is the temporal sampling interval.

Generally speaking, DVS has a high dynamic sensing ability for moving objects in challenging illumination, but it fails to capture static features. Vidar has the ability of high-speed visual texture sampling, but its dynamic range is not as high as that of DVS. Thus, *this paper investigates how to take complementary advantages from Vidar and DVS serve for object detection, especially in challenging scenarios.*

Dataset	Pub.	Resolution	Modality	Type	High-speed	Low-light	Freq.	Classes	Boxes
KITTI (Geiger, Lenz, and Urtasun 2012)	CVPR	1,240×376	Frames	Real	✗	✗	10	3	127.6k
MS COCO (Lin et al. 2014)	ECCV	Various	Frames	Real	✗	✗	1	80	2.5M
KAIST (Hwang et al. 2015)	CVPR	640×480 (RGB) 320×256 (T)	RGB-T	Real	✗	✓	1	1	29k
NightSurveillance (Wang et al. 2020a)	IJCAI	1,920×1,080	Frames	Real	✗	✓	1	1	52k
NFS-CeleX (Huang et al. 2018)	TCSVT	Various	Events	Simulated	✓	✗	240	n.a.	38.3k
PKU-DDD17-CAR (Li et al. 2019)	ICME	346×260	Events, Frames	Real	✗	✓	1	1	3155
Gen1 Detection (de Tournemire et al. 2020)	NeuIPS	304×240	Events	Real	✗	✓	1, 4	2	255k
1 Mpx Detection (Perot et al. 2020)	NeuIPS	1,280×720	Events	Real	✗	✓	60	3	25M
PKU-Vidar-DVS	**Ours**	400×250 (Vidar) 346×260 (DVS)	**Spikes Events**	**Real**	✓	✓	**50**	**9**	**215.5k**

Table 1: Comparison of neuromorphic datasets with bounding boxes and related conventional object detection datasets. Pub. denotes the source of publication. Freq. refers to the frequency of provided labeled bounding boxes.

Multimodal Neuromorphic Dataset

In this section, we first present the details of how to build our PKU-Vidar-DVS dataset. Then, we give detailed statistics to better understand this new dataset.

Collection Setups and Calibration. To test our retinomorphic object detector in real-world scenarios, we collocate a time-based camera (i.e., Vidar, resolution 400×250) and an event-based camera (i.e., DAVIS346, resolution 346×260, including DVS events and RGB frames with 25 FPS). As illustrated in Fig. 2(a), the input light is equally divided into Vidar and DAVIS346 via a beam splitter (i.e., Thorlabs CCM1-BS013) (Wang et al. 2020c; Zhu et al. 2021; Xiang et al. 2021). On this basis, we design the spatio-temporal calibration procedures to synchronize two cameras within the shared view at the same time in Fig. 2(b).

Data Recordings and Annotation. Our PKU-Vidar-DVS dataset contains 9 indoor and outdoor challenging scenarios (see Fig. 2(c)) by considering velocity distribution, illumination change, category diversity, and object scale, etc. We use the hybrid camera system to record 530 sequences including Vidar spikes, DVS events, and RGB frames. In each sequence, we collect approximately 5 seconds as the raw data pool. In order to provide bounding boxes from asynchronous visual streams, frames are reconstructed from Vidar spikes at 50 FPS. After spatio-temporal calibration, all labels are provided by a well-trained professional annotation team.

Data Statistics. Manual annotations in the recordings are provided at a frequency of 50 Hz. As a result, our dataset has 99.6k labeled timestamps and 215.5k labels in total. Afterward, we split them into 165.5k for training, 25k for validation, and 25k for testing. We compare our PKU-Vidar-DVS with the related datasets in Table 1. Notably, this is the first work to build a neuromorphic multimodal object detection dataset involving high-speed and low-light scenes.

Overall, such novel neuromorphic cameras enable our large-scale PKU-Vidar-DVS dataset to be a competitive object detection dataset in challenging scenarios with ***multiple characteristics****: (i) Ultra-high-speed sampling with 12 Meps for DVS and 20,000 Hz spike plane for Vidar, (ii) HDR property with 120 dB for DVS, (iii) Temporally long-term continuous streams with labels at high frequency, (iv) Real-world scenes with abundant diversities in categories and scales.*

Retinomorphic Object Detection

Overview

Given the spatio-temporal window Γ_1, spike streams from Vidar S_1=$\{x_n, y_n, t_n \in \Gamma_1 : n = 1, ..., N_1\}$ are discrete and sparse point sets in 3D space. Similarly, DVS events can be described as S_2=$\{x_n, y_n, t_n, p_n \in \Gamma_2 : n = 1, ..., N_2\}$. In this work, our goal is to accurately detect and identify spatio-temporal objects from asynchronous visual streams, called *retinomorphic object detection* as follows:

Definition 1 (**Retinomorphic Object Detection**). *Let S_1 and S_2 be two asynchronous visual streams from Vidar and DVS respectively, which can be divided into temporal bins S_1=$\left\{S_1^1, S_1^2, ..., S_1^N\right\}$ and S_2=$\left\{S_2^1, S_2^2, ..., S_2^N\right\}$. The corresponding object information B=$\left\{B^1, B^2, ..., B^N\right\}$ can be calculated by:*

$$B^i = \mathcal{D}\left(\left\{S_1^{i-k}, ..., S_1^i\right\}, \left\{S_2^{i-k}, ..., S_2^i\right\}\right), \tag{3}$$

where each B^i=$\{(x_{i,j}, y_{i,j}, w_{i,j}, h_{i,j}, c_{i,j}, t_{i,j})\}_{j\in[1,J]}$ is a list of J bounding box locations and class predictions corresponding to temporal bins S_1^i and S_2^i, where $[x_{i,j}, y_{i,j}]$ are the spatial coordinates of the top left of the j bounding box, $[w_{i,j}, h_{i,j}]$ are its width and height, $c_{i,j}$ and $t_{i,j}$ are the object class and the timestamp. A function $\mathcal{D}$ refers to retinomorphic object detector, and the number of a subset of multiple temporal bins from each stream is $k + 1$.

Towards this end, we propose a novel bio-inspired unifying object detection framework by integrating Vidar and DVS, which aims at addressing the shortages (e.g., motion blur and low-light) of conventional cameras by taking complementary advantages from asynchronous visual streams. As shown in Fig. 3, our retinomorphic object detector consists of *temporal aggregation representation module* and *dynamic interaction fusion module*.

Temporal Aggregation Representation

To make asynchronous streams in combination with deep networks, it is necessary to convert sparse point sets into successive measurements (Fu et al. 2019; Li et al. 2021). Formally, this mapping $\mathcal{M} : S \mapsto E$ is termed as *neuromorphic representation* by:

$$E = \sum_{S\in\Gamma} S\left(x_n, y_n, t_n\right) k\left(x - x_n, y - y_n, t - t_n\right), \tag{4}$$

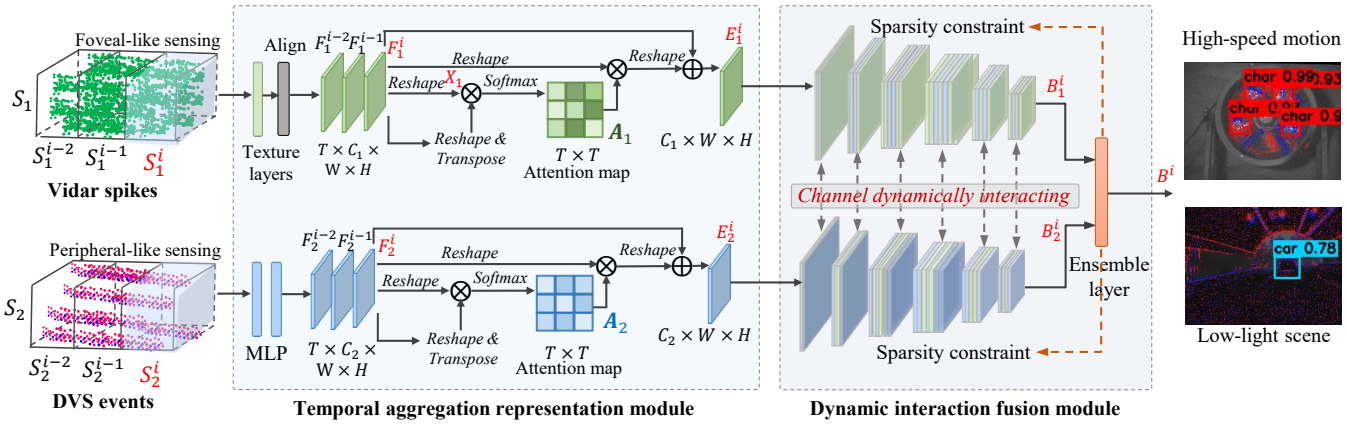


Figure 3: The pipeline of *retinomorphic object detector*. Initially, asynchronous visual streams are split into temporal bins as basic processing units. Then, we design a temporal aggregation representation module using an attention mechanism. Finally, a dynamic interaction fusion module is proposed to combine two streams via exchanging channels between two sub-networks.

where $k\,(x, y, t)$ can adopt handcrafted kernel functions or neural network architecture, and $E(x, y, t)$ should ideally preserve the spatio-temporal information from S.

For Vidar spikes, we first use inter-spike interval (Zhu et al. 2020b) and aligned operation to reconstruct visual textures $F_1^i \in \mathbb{R}^{C_1 \times W \times H}$ with the same resolution of DVS as:

$$F_1^i(x_n, y_n, t_n) = \phi(\frac{d}{t_n - t_{n-1}}) \cdot \boldsymbol{R},\ x_n < W, y_n < H, \quad (5)$$

where d controls the dynamic range of reconstructed textures, ϕ is the non-linearity using gamma correction, and $\boldsymbol{R}$ is a transformation matrix to align reconstructed textures.

For DVS events, we first utilize a **m**ulti-**l**ayer **p**erceptron (MLP) (Gehrig et al. 2019) with two hidden layers to generate features $F_2^i \in \mathbb{R}^{C_2 \times W \times H}$ for each event sequence as:

$$F_2^i(x_n, y_n, t_n) = \sigma\left(\sum \omega\,(t_n^*) \cdot p_n\right), x_n < W, y_n < H, \quad (6)$$

where the weight $\omega\,(t_n^*)$ is a learning parameter, the activation function $\sigma\,(\cdot)$ is chosen ReLU, $t_n^* = \frac{t_n - t_1}{\Delta t}$ is the normalized timestamp, and $p_n \in \{1, -1\}$ refers to the polarity.

Take temporal feature aggregation of Vidar for example, we first compute the temporal attention map $A_1 \in \mathbb{R}^{T \times T}$ from the original Vidar features $\left\{F_1^{i-T+1}, ..., F_1^i\right\} \in \mathbb{R}^{T \times C_1 \times W \times H}$. Specifically, we reshape Vidar features to $X_1 \in \mathbb{R}^{T \times N}$, and then perform a matrix multiplication between X_1 and the transpose of X_1 (see Fig. 3). Finally, we use a softmax layer to calculate the temporal attention map $A_1 \in \mathbb{R}^{T \times T}$ as follows:

$$a_1^{ij} = \frac{\exp(X_1^i \cdot X_1^j)}{\sum_{j=1}^{T} \exp(X_1^i \cdot X_1^j)}, \quad (7)$$

where a_1^{ij} denotes the impact of i^{th} and j^{th} feature maps, and T is the temporal aggregation size.

The temporal aggregation representation E_1^i for multiple adjacent temporal bins can be described as follows:

$$E_1^i = \sum_{j=1}^{T} a_1^{ij} X_1^j + F_1^i. \quad (8)$$

Dynamic Interaction Fusion

The goal of deep multimodal fusion is to make the output B of the detector $\mathcal{D}$ to fit the label $\bar{B}$ as much as possible, and it can formulate the following minimization problem as:

$$\min_{\mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(B^i = \mathcal{D}(F^i), \bar{B}^i), \quad (9)$$

where $F^i = \left\{F_m^i\right\}_{m=1}^{M}$ refers to the i_{th} feature map from M modalities, N is the batch size, and $\mathcal{L}$ is the loss function.

Inspired by the dynamic interaction between foveal and peripheral signals (Stewart, Valsecchi, and Schütz 2020), we introduce a dynamic interaction fusion module via exchanging the channels (Wang et al. 2020b), instead of typical aggregation fusion operations (e.g., averaging (Li, Wu, and Kittler 2020) and concatenation (Xu et al. 2020a)). The whole optimization objective function can be formulated as:

$$\min_{\mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\sum_{m=1}^{M} \alpha_m \mathcal{D}_m(F_m^i), \bar{B}^i) + \lambda \sum_{m=1}^{M} \sum_{l=1}^{L} |\gamma_{m,l}|\,, \quad (10)$$

where the decision score α_m determines the final result in an ensemble layer by training the softmax. The scaling factors $\gamma_{m,l}$ for the l-th layer channel in the m-th modality are computed using Batch Normalization (BN). We adopt the ℓ_1 norm to penalize the scaling factors for channel sparsity. λ is a hyperparameter that weights the relative contribution of the norm penalty term to avoid overfitting.

In fact, the scaling factor of the BN layer reflects the importance of the feature map in the c-th channel, which is replaced by the mean $F'_{m,l,c}$ of other modalities once the scaling factor is smaller than a presetting threshold θ_γ as:

$$F'_{m,l,c} = \frac{1}{M} \sum_{m' \neq m}^{M} \gamma_{m',l,c} \frac{F_{m',l,c} - \mu_{m',l,c}}{\sqrt{\sigma_{m',l,c}^2 + \varepsilon}} + \beta_{m',l,c}, \quad (11)$$

where $F_{m',l,c}$ is the c-th channel before the l-th BN layer in the m'-th sub-network. ε is a small constant, and $\mu_{m',l,c}$ and $\sigma_{m',l,c}$ denote the mean and the standard deviation. $\gamma_{m',l,c}$ and $\beta_{m',l,c}$ are the trainable scaling factor and the offset.

Modality	Method	Representation	Fusion	Backbone	PKU-Vidar-DVS		KITTI simulated	
					mAP	Runtime (ms)	mAP	Runtime (ms)
DVS	(Chen 2018)	Event image	-	YOLO	0.331	12.26	0.307	19.34
	(Iacono et al. 2018)	Event image	-	SSD	0.326	10.61	0.301	17.80
	NGA-events (Hu et al. 2020)	Event volume	-	YOLOv3	0.353	15.37	0.349	20.97
	Our baseline	E2vid	-	YOLOv3	0.394	197.81	0.371	235.62
		TAR-events	-	YOLOv3	<u>0.386</u>	16.49	<u>0.356</u>	21.33
Vidar	Our baseline	Spike image	-	YOLOV3	0.497	13.70	0.612	19.86
		VTTW	-	YOLOv3	0.516	13.70	0.644	19.86
		VTII	-	YOLOv3	0.551	13.70	0.673	19.86
		TAR-spikes	-	YOLOv3	**0.579**	<u>17.34</u>	**0.701**	<u>22.47</u>
Vidar+DVS	(Jiang et al. 2019)	Event image+VTII	NMS	YOLOv3	0.586	17.91	0.713	23.02
	JDF (Li et al. 2019)	Event image+VTII	Score fusion	YOLOv3	0.591	17.93	0.718	23.11
	Ours	**TAR**	**Dynamic interaction**	YOLOv3	**0.647**	19.05	**0.762**	25.35

Table 2: Performance evaluation of our PKU-Vidar-DVS dataset and KITTI simulated dataset. Our retinomorphic object detector, integrating Vidar and DVS by **t**emporal **a**ggregation **r**epresentation (TAR) and dynamic interaction fusion, outperforms five state-of-the-art methods and our six baselines, especially the single-modality.

Experiments

This section will first describe the experimental settings. Then, we conduct the effective test and ablation test to verify our approach. Finally, the scalability test provides quantitative results in various motion speeds and light conditions.

Experimental Settings

Datasets. To verify the effectiveness of our retinomorphic object detector, we conduct experiments on our newly built PKU-Vidar-DVS dataset and KITTI simulated dataset (Geiger, Lenz, and Urtasun 2012). More precisely, KITTI simulated dataset, including 20 videos for object tracking, is converted from videos to Vidar spikes using our open-source Vidar simulator (Kang et al. 2021) and DVS events by V2E simulator (Hu, Liu, and Delbruck 2021). This simulated dataset consists of 14 asynchronous visual streams for training, 3 hybrid streams (i.e., 0007, 0017, and 0018) for validating, and the remaining 3 hybrid streams (i.e., 0000, 0003, and 0006) for testing. Besides, the light degradation ratio η=2 is set to simulate low-light scenarios for our Vidar simulator (i.e., linear light sensing) and the V2E simulator (i.e., logarithmic light sensing).

Implementation Details. We set the overlap threshold to 0.5, the predicting score to 0.5 for Vidar, and 0.3 for DVS. Two widely used metrics (i.e., COCO mAP (Lin et al. 2014) and runtime (ms)) in object detection tasks are adopted to report the performance scores. In the fusion stage, we select YOLOv3 (Redmon and Farhadi 2018) as a fair sub-network considering the balance of the accuracy and time complexity, and two branches are dynamically exchanging channels simultaneously. All networks are trained for 60 epochs with the Adam optimizer on an NVIDIA Tesla V100-PCLE GPU with the learning rate of 10^{-4}. We set λ to 10^{-3} for sparsity constrains in Equation (10) and the threshold θ_γ to 10^{-2} for dynamic interaction fusion in Equation (11). We set the temporal aggregation size T as 3 to make an accuracy-speed trade-off. We utilize the best training model on the validation dataset and apply it to the testing dataset to report the final detection performance.

Effective Test

We will investigate that why and how our retinomorphic object detector works from the following three perspectives.

Evaluation on DVS Modality. To evaluate our **t**emporal **a**ggregation **r**epresentation for DVS **events** (TAR-events), we compare TAR-events with other representations from three event-based object detectors (i.e., event images for YOLO (Redmon et al. 2016), event images for SSD (Liu et al. 2016), and event volumes for YOLOv3) and our another baseline (i.e., reconstructing video using E2vid (Rebecq et al. 2019) for YOLOv3). As illustrated in Table 2, our baseline, using TAR-events for YOLOv3, obtains better performance than these three existing methods meanwhile maintaining comparable computational speed. Our other baseline, utilizing E2vid to generate gray frames, can achieve the best performance, but the two stages of first image reconstruction and then object detection are very complicated and time-consuming.

Evaluation on Vidar Modality. We present four representations for Vidar spikes as our baselines including: (i) mapping Vidar stream into spike images, (ii) **v**isual **t**exture from **t**emporal **w**indow (VTTW), (iii) **v**isual **t**exture from **i**nterspike **i**nterval (VTII) (Dong et al. 2019; Zhu et al. 2019, 2020b,a), (iv) **t**emporal **a**ggregation **r**epresentation for Vidar **spikes** (TAR-spikes). From Table 2, our TAR-spikes has the improvement over three strategies (i.e., spike image, VTTW, and VTII) by a large margin, with an average of 5.7% and 5.8% increase in mAP for our PKU-Vidar-DVS dataset and KITTI simulated dataset respectively. The improvement is because that our TAR-spikes can produce fine-tuned textures from Vidar spikes and leverage rich temporal information from adjacent temporal bins. Besides, the computational speed of our TAR-spikes, introducing temporal aggregation strategy, is almost comparable in contrast to other strategies without using temporal cues.

Benefit From Dynamic Interaction Fusion. To make a comparison with joint detection frameworks (Jiang et al. 2019; Li et al. 2019) as fair as possible, we use VTII from Vidar instead of RGB frames from DAVIS346. As depicted

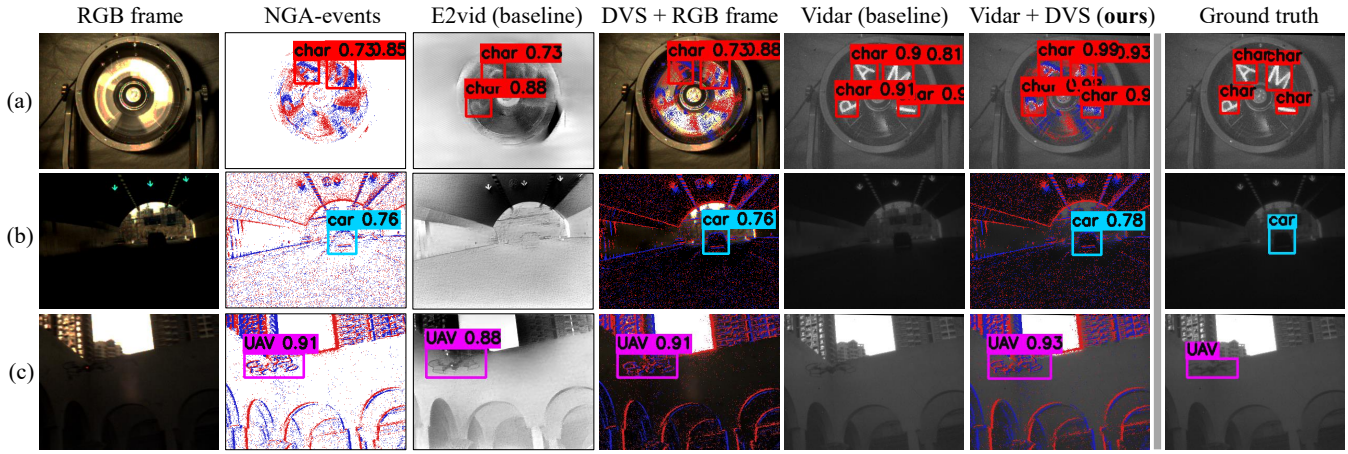


Figure 4: Representative visualization results on our PKU-Vidar-DVS dataset. (a) High-speed rotation characters with 1500 r/min. (b) Rushing car in a tunnel. (c) High-speed moving UAV with low-light. Note that, conventional frames from DAVIS346 suffer from motion blur and insufficient texture. On the contrary, our retinomorphic detector, inheriting HDR property from DVS and high-speed visual texture from Vidar, outperforms the single-modality methods in high-speed and low-light scenes.

Method	The baseline	(a)	(b)	(c)	**Ours**
TAR-spikes		✓		✓	✓
TAR-events			✓	✓	✓
Dynamic interaction			✓	✓	✓
ℓ_1 regulation					✓
mAP	0.551	0.579	0.606	0.623	**0.647**
Runtime (ms)	13.70	17.34	17.85	18.99	19.05

Table 3: Performance components of our approach on PKU-Vidar-DVS dataset. All results are obtained with our baseline using VTII from Vidar and YOLOv3.

in Table 2, our retinomorphic object detector, incorporating Vidar spikes and DVS events by temporal aggregation representation and dynamic interaction fusion, achieves the best performance in contrast to five state-of-the-art methods and our six baselines. More precisely, our approach has a 9.6% mAP improvement over the baseline (i.e., VTII from Vidar) by bringing DVS events, this is because DVS with the HDR property might be more significant when Vidar stream is affected in low-light scenes. Comparing with a single DVS modality, our method truly shines, outperforming the baseline (i.e., TAR-events) on our PKU-Vidar-DVS dataset by a large margin (0.647 versus 0.386), this indicates that Vidar with high-speed visual textures may be more important for high-precision recognition. We claim that our approach has comparable time complexity with the single-modality.

We further present representative visualization results on our PKU-Vidar-DVS dataset (see Fig. 4). Apparently, our retinomorphic object detector outperforms single-modality methods in high-speed and low-light scenarios. On the contrary, RGB frames fail to detect objects in these scenarios. We find that DVS, offering high temporal resolution and HDR, has provided insight into overcoming the shortages of conventional cameras, but it is hard to achieve high-precision recognition due to the lack of rich texture

Method	PKU-Vidar-DVS		KITTI simulated	
	mAP	Runtime (ms)	mAP	Runtime (ms)
NMS	0.612	18.02	0.735	23.51
Score fusion	0.615	18.11	0.739	23.80
Averaging (early)	0.619	18.25	0.742	24.01
Averaging (middle)	0.623	18.30	0.746	24.22
Averaging (late)	0.621	18.33	0.744	24.28
Concatenation (early)	0.624	18.64	0.748	24.67
Concatenation (middle)	0.627	18.72	0.752	24.73
Concatenation (late)	0.626	18.79	0.750	24.80
Interaction (Ours)	**0.647**	19.05	**0.762**	25.35

Table 4: Comparison with typical fusion methods including the post-processing (e.g., NMS and score fusion) and feature aggregation operations (e.g., averaging and concatenation).

in Fig. 4(a). Unfortunately, the joint framework using DVS events and RGB frames still remains a bottleneck in the limited frame rate. From Fig. 4(b)-(c), Vidar is difficult to capture objects in low-light scenes.

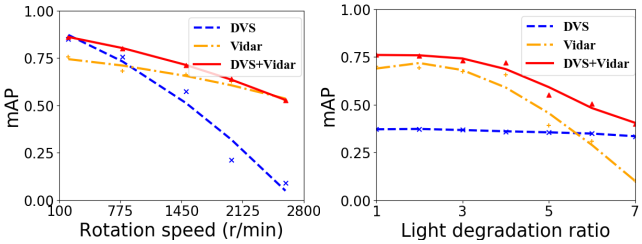
Ablation Test

Beyond effective tests, we next conduct ablation tests to take a deep look at the impact of each design choice as follows.

Contribution of Each Component. As shown in Table 3, three methods, namely (a)-(c), utilize TAR-spikes from Vidar, dynamic interaction for two sub-networks, and ℓ_1 regulation for channel sparsity respectively, consistently achieve higher performance on PKU-Vidar-DVS dataset than the baseline using VTII from Vidar and YOLOv3. Specifically, our TAR-spikes, leveraging rich temporal cues, obtains the 2.8% mAP improvement over the baseline. Comparing (a) and the baseline, the absolute promotion is 5.5%, which demonstrates that it is feasible to adopt dynamic interaction fusion between two sub-networks. Furthermore, we improve the mAP from 0.623 to 0.647, where the only difference be-

Dataset	Size T	1	3	5	7	9
PKU-Vidar-DVS	mAP	0.623	**0.647**	0.649	0.650	0.650
	Runtime (ms)	18.17	19.05	19.78	20.94	22.72
KITTI simulated	mAP	0.738	**0.762**	0.768	0.769	0.768
	Runtime (ms)	24.14	25.35	25.97	27.31	28.49

Table 5: Detection performance with different temporal aggregation sizes from asynchronous visual streams.



(a) Motion speed (b) Light condition

Figure 5: Quantitative evaluation. (a) Rotation characters with various speeds on PKU-Vidar-DVS dataset. (b) Simulating different light intensities on KITTI simulated dataset.

tween them is whether using ℓ_1 regulation for channel sparsity. And the last row of Table 3 illustrates that the computational speeds of these methods are almost comparable.

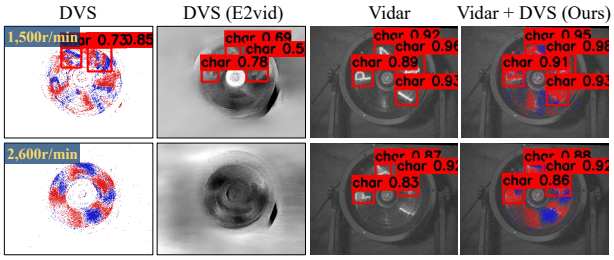
Comparison with Typical Fusion Methods. As illustrated in Table 4, our strategy achieves the best performance against the post-posting (e.g., NMS (Jiang et al. 2019) and score fusion (Li et al. 2019)) and end-to-end feature aggregation operations (e.g., averaging (Li, Wu, and Kittler 2020) and concatenation (Xu et al. 2020a; Wang et al. 2021c,a)) with three fusion stages. For example, our approach gets around 3.5%, 3.2%, 2.6% and 2.1% improvements on our PKU-Vidar-DVS dataset with typical fusion methods meanwhile keeps almost comparable computational cost.

Temporal Aggregation Size. We test temporal aggregation module with different aggregation sizes (e.g., T=3, 5, 7, and 9) in Table 5. For instance, the corresponding mAPs on our PKU-Vidar-DVS dataset improve 2.4%, 2.6%, 2.7%, and 2.7% respectively. Obviously, as we amplify the temporal aggregation size, the detection accuracy keeps improving meanwhile increasing the computational speed. Thus, we set the temporal aggregation size as T=3 for a good balance.

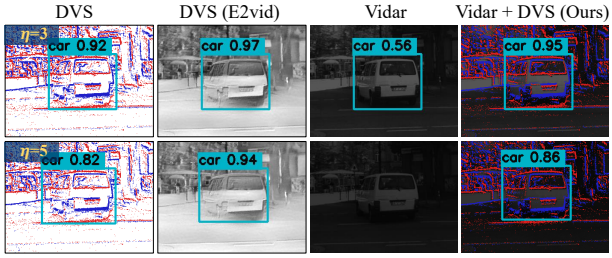
Scalability Test

To investigate the properties of Vidar and DVS, we present quantitative results on various speeds and light intensities.

Quantitative Evaluation for High-Speed Motions. We design a specific scenario (i.e., moving characters with various rotation speeds) to verify the robustness of our retinomorphic object detector. We set five speed levels (i.e., 200, 800, 1,500, 2,000, and 2,600 r/min) to record rotation characters on an electric fan. As shown in Fig. 5(a), our approach, integrating Vidar and DVS, has better detection performance than the single-modality. Besides, the blue curve using DVS drops sharply along with the increase of rotation speeds,



(a) Rotation characters on a fan from our PKU-Vidar-DVS dataset



(b) Driving car on KITTI simulated dataset (clipped into 346×240)

Figure 6: Visualization results under different motion speeds and light intensities. (a) Recording high-speed motion characters with 1,500 and 2,600 r/min. (b) Simulating low-light scenes with light degradation ratios η =3 and η =5.

but the other two curves involving Vidar decrease gradually. This may be caused by the fact that Vidar has a strong ability to capture high-speed moving objects (see Fig. 6(a)).

Quantitative Evaluation for Light Changes. We set light degradation ratios (i.e., from η =2 to 7) to reduce the light intensity for the integrator in our Vidar simulator (i.e., linear light sensing) and the comparator in the V2E simulator (i.e, logarithmic light sensing). Some visualization results for the KITTI simulated dataset are reported in Fig. 6(b), our approach performs well in low-light scenarios. The mAP of Vidar decreases sharply with an increase in degradation ratio of η in Fig. 5(b), but the curve of DVS remains relatively stable. In other words, after incorporating the auxiliary DVS streams, our approach improves significantly the performance over only using Vidar in low-light scenarios.

Conclusion

This paper presents a novel *retinomorphic object detector* to overcome common object challenges (e.g., motion blur and low-light). To the best of our knowledge, this is the first work to explore such a novel object detector integrating foveal-like and peripheral-like sensing as the retina does. Moreover, we develop a hybrid camera system, build a large-scale multimodal neuromorphic object detection dataset (i.e., PKU-Vidar-DVS), and design an open-source Vidar simulator. The results show that our approach outperforms the state-of-the-art methods within the single-modality, which inherits high-speed visual textures from Vidar and the HDR property from DVS. We believe this work will be the key to taking the advantage of neuromorphic cameras on various vision tasks in challenging scenes. We also believe this prototype will provide insight into next-generation neuromorphic cameras.

Acknowledgment

This work is partially supported by grants from the National Natural Science Foundation of China under contract No. 62027804, No. 61825101, No. 62088102, No. 62132002, and No. 6210225, the Postdoctoral Innovative Talent Support Program BX20200174, and the China Postdoctoral Science Foundation Funded Project 2020M682828.

References

Boahen, K. A. 2000. Point-to-point connectivity between neuromorphic chips using address events. *IEEE Trans. Circuits Syst., II, Exp. Briefs*, 47(5): 416–434.

Brandli, C.; Berner, R.; Yang, M.; Liu, S.-C.; and Delbruck, T. 2014. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE J. Solid-State Circuits*, 49(10): 2333–2341.

Cao, H.; Chen, G.; Xia, J.; Zhuang, G.; and Knoll, A. 2021. Fusion-based feature attention gate component for vehicle detection based on event camera. *IEEE Sensors J.*, 21(21): 24540–24548.

Chen, D. G.; Matolin, D.; Bermak, A.; and Posch, C. 2011. Pulse-modulation imaging—Review and performance analysis. *IEEE J. Solid-State Circuits*, 5(1): 64–82.

Chen, N. F. 2018. Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion. In *CVPRW*, 644–653.

de Tournemire, P.; Nitti, D.; Perot, E.; Migliore, D.; and Sironi, A. 2020. A large scale event-based detection dataset for automotive. In *arXiv*, 1–8.

Dong, S.; Huang, T.; and Tian, Y. 2017. Spike camera and its coding methods. In *DCC*, 437–437.

Dong, S.; Zhu, L.; Xu, D.; Tian, Y.; and Huang, T. 2019. An efficient coding method for spike camera using inter-spike intervals. In *DCC*, 568–568.

Fu, Y.; Li, J.; Dong, S.; Tian, Y.; and Huang, T. 2019. Spike coding: Towards lossy compression for dynamic vision sensor. In *DCC*, 572–572.

Gallego, G.; Delbruck, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*

Gehrig, D.; Loquercio, A.; Derpanis, K. G.; and Scaramuzza, D. 2019. End-to-end learning of representations for asynchronous event-based data. In *ICCV*, 5633–5643.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The kitti vision benchmark suite. In *CVPR*, 3354–3361.

Hu, Y.; Delbruck, T.; Liu, S.-C.; and Liu, S.-C. 2020. Learning to exploit multiple vision modalities by using grafted networks. In *ECCV*, 85–101.

Hu, Y.; Liu, S.-C.; and Delbruck, T. 2021. V2e: From video frames to realistic DVS events. In *CVPRW*, 1312–1321.

Huang, J.; Wang, S.; Guo, M.; and Chen, S. 2018. Event-guided structured output tracking of fast-moving objects using a CeleX sensor. *IEEE Trans. Circuits Syst. Video Technol.*, 28(9): 2413–2417.

Hwang, S.; Park, J.; Kim, N.; Choi, Y.; and So Kweon, I. 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 1037–1045.

Iacono, M.; Weber, S.; Glover, A.; and Bartolozzi, C. 2018. Towards event-driven object detection with off-the-shelf deep learning. In *IROS*, 1–9.

Jiang, Z.; Xia, P.; Huang, K.; Stechele, W.; Chen, G.; Bing, Z.; and Knoll, A. 2019. Mixed frame-/event-driven fast pedestrian detection. In *ICRA*, 8332–8338.

Kang, Z.; Li, J.; Zhu, L.; and Tian, Y. 2021. Retinomorphic sensing: A novel paradigm for future multimedia computing. In *ACM MM*, 144–152.

Katzakis, N.; Chen, L.; Teather, R. J.; Ariza, O.; and Steinicke, F. 2019. Evaluation of 3D pointing accuracy in the fovea and periphery in immersive head-mounted display environments. *IEEE Trans. Vis. Comput. Graphics*.

Li, H.; Wu, X.-J.; and Kittler, J. 2020. MDLatLRR: A novel decomposition method for infrared and visible image fusion. *IEEE Trans. Image Process.*, 29: 4733–4746.

Li, J.; Dong, S.; Yu, Z.; Tian, Y.; and Huang, T. 2019. Event-based vision enhanced: A joint detection framework in autonomous driving. In *ICME*, 1396–1401.

Li, J.; Fu, Y.; Dong, S.; Yu, Z.; Huang, T.; and Tian, Y. 2021. Asynchronous spatiotemporal spike metric for event cameras. *IEEE Trans. Neural Netw. Learn. Syst.*

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008. A 128×128 120 dB 15μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2): 566–576.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Liu, M.; Qi, N.; Shi, Y.; and Yin, B. 2021. An attention fusion network for event-based vehicle object detection. In *ICIP*, 3363–3367.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single shot multibox detector. In *ECCV*, 21–37.

Perot, E.; de Tournemire, P.; Nitti, D.; Masci, J.; and Sironi, A. 2020. Learning to detect objects with a 1 megapixel event camera. In *NeuIPS*, 16639–16652.

Posch, C.; Serrano-Gotarredona, T.; Linares-Barranco, B.; and Delbruck, T. 2014. Retinomorphic event-based vision sensors: bioinspired cameras with spiking output. *Proc. IEEE*, 102(10): 1470–1484.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, 3857–3866.

Redmon, J.; Divvala, S. K.; Girshick, R. B.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*, 779–788.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. In *arXiv*, 1–6.

Roy, K.; Jaiswal, A.; and Panda, P. 2019. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784): 607–617.

Ryan, C.; O'Sullivan, B.; Elrasad, A.; Cahill, A.; Lemley, J.; Kielty, P.; Posch, C.; and Perot, E. 2021. Real-time face & eye tracking and blink detection using event cameras. *Neural Netw.*, 141: 87–97.

Sayed, M.; and Brostow, G. 2021. Improved handling of motion blur in online object detection. In *CVPR*, 1706–1716.

Sinha, R.; Hoon, M.; Baudin, J.; Okawa, H.; Wong, R. O.; and Rieke, F. 2017. Cellular and circuit mechanisms shaping the perceptual properties of the primate fovea. *Cell*, 168(3): 413–426.

Stewart, E. E.; Valsecchi, M.; and Schütz, A. C. 2020. A review of interactions between peripheral and foveal vision. *J. Vis.*, 20(12): 2–2.

Strasburger, H.; Rentschler, I.; and Jüttner, M. 2011. Peripheral vision and pattern recognition: A review. *J. Vis.*, 11(5): 13–13.

Wang, X.; Chen, J.; Wang, Z.; Liu, W.; Satoh, S.; Liang, C.; and Lin, C.-W. 2020a. When pedestrian detection meets nighttime surveillance: A new benchmark. In *IJCAI*, 509–515.

Wang, X.; Chen, Z.; Tang, J.; Luo, B.; Wang, Y.; Tian, Y.; and Wu, F. 2021a. Dynamic attention guided multi-trajectory analysis for single object tracking. *IEEE Trans. Circuits Syst. Video Technol.*

Wang, X.; Li, J.; Zhu, L.; Zhang, Z.; Chen, Z.; Li, X.; Wang, Y.; Tian, Y.; and Wu, F. 2021b. VisEvent: Reliable Object Tracking via Collaboration of Frame and Event Flows. In *arXiv*.

Wang, X.; Shu, X.; Zhang, Z.; Jiang, B.; Wang, Y.; Tian, Y.; and Wu, F. 2021c. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *CVPR*, 13763–13773.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020b. Deep multimodal fusion by channel exchanging. *NeuIPS*, 33.

Wang, Z. W.; Duan, P.; Cossairt, O.; Katsaggelos, A.; Huang, T.; and Shi, B. 2020c. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *CVPR*, 1609–1619.

Wu, Y.; Deng, L.; Li, G.; Zhu, J.; Xie, Y.; and Shi, L. 2019. Direct training for spiking neural networks: Faster, larger, better. In *AAAI*, 1311–1318.

Xiang, X.; Zhu, L.; Li, J.; Wang, Y.; Huang, T.; and Tian, Y. 2021. Learning super-resolution reconstruction for high temporal resolution spike stream. *IEEE Trans. Circuits Syst. Video Technol.*

Xu, J.; Wang, W.; Wang, H.; and Guo, J. 2020a. Multi-model ensemble with rich spatial information for object detection. *Pattern Recognit.*, 99: 107098.

Xu, J.; Xu, L.; Gao, Z.; Lin, P.; and Nie, K. 2020b. A denoising method based on pulse interval compensation for high-speed spike-based image sensor. *IEEE Trans. Circuits Syst. Video Technol.*

Zheng, H.; Wu, Y.; Deng, L.; Hu, Y.; and Li, G. 2021. Going deeper with directly-trained larger spiking neural networks. In *AAAI*, 11062–11070.

Zhu, L.; Dong, S.; Huang, T.; and Tian, Y. 2019. A retina-inspired sampling method for visual texture reconstruction. In *ICME*, 1432–1437.

Zhu, L.; Dong, S.; Huang, T.; and Tian, Y. 2020a. Hybrid Coding of Spatiotemporal Spike Data for a Bio-inspired Camera. *IEEE Trans. Circuits Syst. Video Technol.*

Zhu, L.; Dong, S.; Li, J.; Huang, T.; and Tian, Y. 2020b. Retina-like visual image reconstruction via spiking neural model. In *CVPR*, 1438–1446.

Zhu, L.; Li, J.; Wang, X.; Huang, T.; and Tian, Y. 2021. NeuSpike-Net: High speed video reconstruction via bio-inspired neuromorphic cameras. In *ICCV*, 2400–2409.