eTraM: Event-based Traffic Monitoring for Resource-Efficient Detection and Tracking Across Varied Lighting Conditions

by

Aayush Atul Verma

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2024 by the
Graduate Supervisory Committee:

Yezhou Yang, Chair
Hua Wei
Bharatesh Chakravarthi
Duo Lu

ARIZONA STATE UNIVERSITY

May 2024

ABSTRACT

Traffic monitoring plays a crucial role in urban planning, transportation management, and road safety initiatives. However, existing monitoring systems often struggle to balance the need for high-resolution data acquisition and resource efficiency. This study proposes an innovative approach leveraging neuromorphic sensor technology to enhance traffic monitoring efficiency while still exhibiting robust performance when exposed to difficult conditions. Neuromorphic cameras, also called event-based cameras, with their high temporal and dynamic range and minimal memory usage, have found applications in various fields. However, despite their potential, their use in static traffic monitoring is largely unexplored. This study introduces eTraM, the first-of-its-kind fully event-based traffic monitoring dataset, to address the gap in existing research. eTraM offers 10 hr of data from diverse traffic scenarios under varying lighting and weather conditions, providing a comprehensive overview of real-world situations. Providing 2M bounding box annotations, it covers eight distinct classes of traffic participants, ranging from vehicles to pedestrians and micro-mobility. eTraM's utility has been assessed using state-of-the-art methods, including RVT, RED, and YOLOv8. The quantitative evaluation of the ability of event-based models to generalize on nighttime and unseen scenes further substantiates the compelling potential of leveraging event cameras for traffic monitoring, opening new avenues for research and application.

DEDICATION

Dedicated to my loving parents, family, and friends, whose unwavering love, patience, and faith have been the bedrock of support throughout this journey and will continue to be in all that lies ahead.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to all those who have contributed to the completion of this thesis on a new research topic. Tackling this unexplored territory has been both challenging and rewarding, significantly enriching my understanding of the subject matter.

I am especially grateful to my advisor, Dr. Yezhou Yang, for his invaluable guidance, unwavering support, and profound insights throughout the entirety of this research endeavor. His input has been pivotal in shaping the direction of this work and refining its content. I extend my special thanks to Dr. Bharatesh Chakravarthi for his constant availability for discussions and assistance in helping navigate the intricacies of the thesis and career-related matters. I also deeply appreciate the input provided by Dr. Hua Wei and Dr. Duo Lu, whose constant feedback, suggestions, and expert advice have significantly strengthened the quality of this study.

I would also like to thank Arpit and Rohith for their invaluable involvement during the brainstorming sessions. Moreover, thanks to Manideep, Keshava, Prithvi, and Jayaram for their dedicated efforts in the annotations.

To my parents, family, and friends, I owe an immeasurable debt of gratitude. Their presence in my life has made this journey not only possible but also immensely fulfilling and exciting.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER

1 INTRODUCTION 1

1.1 Motivation and Contribution 2

2 BACKGROUND 6

2.1 Neuromorphic Sensing 7

2.2 Event-based Datasets 11

2.2.1 Early event-based datasets 11

2.2.2 Ego-motion event-based datasets 12

2.2.3 Fixed perception event-based datasets 13

2.3 Event Representations 14

2.3.1 Individual Events 14

2.3.2 Image/Tensor Representation 15

2.3.3 Time-Surfaces 16

2.3.4 Voxel-based 17

2.3.5 Graph-based 18

3 THE *ETRAM* DATASET 19

3.1 Dataset Acquisition Framework 19

3.2 Preprocessing and Annotation 20

3.3 Dataset Statistics 23

4 BENCHMARKING *ETRAM* 29

4.1 Traffic Participant Detection on *eTraM* 30

4.2 Multi-Object Tracking on *eTraM* 33

CHAPTER Page

4.3 Impact of Object Size on Detection Performance 34
5 GENERALIZATION CAPABILITIES OF EVENT DATA 36
5.1 Generalization on Night time 36
5.2 Generalization on Unseen Scenes 38
6 DISCUSSION 40
6.1 Event Cameras in Traffic Monitoring 40
6.2 Static and Ego Event-based Datasets 42
7 CONCLUSION 44
7.1 Summary 44
7.2 Limitations and Future Work 45
REFERENCES 46

LIST OF TABLES

Table Page

2.1 A Comprehensive Review Of Event-based Traffic Datasets From 2017 To 2024. (VH - Vehicle, PED - Pedestrian, MM - Micro-mobility) 14

3.1 Format Of An Asynchronous Event. 22

3.2 2D Bounding Box Annotation Format In *ETraM*. 22

4.1 Baseline Evaluation: Comprehensive Evaluation Of State-of-the-art Tensor-based Approaches RVT, RED, And Frame-based Approach YOLOv8 Across Various Traffic Sites (Intersections, Roadways, Local Streets) During Both Daytime And Nighttime For PED - Pedestrian, VH - Vehicle, And MM - Micro-mobility. 30

4.2 Evaluation Of Object Size Impact On The Performance Of RVT And RED. 35

5.1 Evaluation Of Generalization Capabilities Of RED And RVT On Night Time Data For PED - Pedestrian And VH - Vehicle Class For Models Trained On Only Daytime And A Combination Of Daytime And Nighttime Data. 37

5.2 Evaluation Of Generalization Capabilities Of RED And RVT On Unseen Traffic Scenarios For PED - Pedestrian And VH - Vehicle Tested On Held In And Held Out Test Set. 39

LIST OF FIGURES

Figure Page

1.1 Unveiling The Dynamic World Of Road Traffic: A Glimpse Into Our Event-based Traffic Monitoring Dataset Featuring Diverse Traffic Participants, Including Pedestrians, Various Sized Vehicles, And Micro-mobility Users That Include Cyclists, Wheelchair Users, And Bikers. .. 5

2.1 Event Generation Model: (a) Is The Compact IMX636HD Event Camera By Prophesee And SONY, And (b) Graphically Illustrates (Prophesee S.A (2024a)) How CD ON (Positive Polarity) And CD OFF (Negative Polarity) Events Are Triggered Due To A Change In The Log Of Photocurrent. 7

2.2 Contrasting Standard Camera With Event Camera Illustrating High Temporal Resolution: (a) Demonstrates The Data Loss In Standard Cameras Between Consecutive Frames (redrawn From Mueggler *et al.* (2014)). Meanwhile, (b) Illustrates Motion Blur Typical In Frame-based Cameras During Fast Motion, A Phenomenon Absent In Event-based Cameras. 10

3.1 Data Collection Setup: The First Four Images From The Top Left Display Daytime Data Collection Sites, The Center Image Shows The Prophesee EVK4 HD Camera And The Last Four Images Depict Nighttime Collection Sites. 20

3.2 Impact Of Spatiotemporal Filtering On Event Camera Data: Comparison Of A Noisy Pre-filtered Image (Left) And The Enhanced Clarity Achieved Post-filtering (Right) On Daytime (Top Row) And Nighttime Data (Bottom Row) 21

Figure Page

3.3 A Histogram Illustrating The Event-time Frequency Of *ETraM* (Static Event Dataset) As Compared To 1 Megapixel And DSEC (Ego-motion Event Datasets). 23

3.4 The Object Density Of Various Classes Across The Frame. 24

3.5 Power-law Distribution Of The Number Of Instances Within An Image For Most Predominant Classes - Cars And Pedestrians. 25

3.6 Distribution Of Two Major Traffic Participant Categories Across Various Traffic Sites. 26

3.7 The Bar Plot Illustrates The Average Duration, In Seconds, Spent By Instances Of Different Classes, Providing Insights Into The Temporal Characteristics Of Each Class In The Dataset. 27

3.8 The Bar Plot Illustrates The Average Duration, In Seconds, Spent By Instances Of Different Classes, Providing Insights Into The Temporal Characteristics Of Each Class In The Dataset. 28

4.1 Traffic Participant Object Detection By RVT. Snapshots Illustrating The Detection Results Of RVT At Various Traffic Sites, Showcasing Its Performance In Diverse Real-world Scenarios. 32

4.2 Traffic Participant Object Detection By RED. Snapshots Illustrate The Detection Results Of RED At Various Traffic Sites, Showcasing Its Performance In Diverse Real-world Scenarios. 32

4.3 Qualitative Results: Showcasing Ground Truth (Top Row) Annotations In *ETraM* And The Corresponding Tracking Of Each Detected Object (Bottom Row) Where Each Trailing Line Denotes The Path Followed By The Detected Object In Previous Timesteps. 34

Figure Page

6.1 Demonstrating Effectiveness Of Event Camera For Traffic Scenarios: Yellow Circle (Top Row) Tracks A Car That Halts At A Stop Sign With Lack Of Motion Captured In The Third Frame, Red Circle (Bottom Row) Tracks A Car That Violates The Stop Sign Where Motion Is Continuously Captured In Every Frame. Additionally, The Green Arrow (Top Row) Shows A Car Traveling At A High Speed, Resulting In A High Event Density. 41

6.2 Qualitative Comparison: Events Captured From A Static Event Camera (Left) Show Enhanced Visibility Of Moving Vehicles Compared To An Ego-motion Event Camera (Right). 42

6.3 Traffic Site Diversity In *ETraM*: Various Instances Encapsulating The Interactions Amongst Multiple Traffic Participants Captured From A Static Roadside Perspective Are Shown With Daytime (First Row), Twilight (Second Row), And Nighttime (Last Row) Showing Increasing Sensor Noise (Top To Bottom) Due To Light Sources Such As Headlights And Streetlights. 43

Chapter 1

INTRODUCTION

Intelligent Transportation Systems (ITS) represent the intersection of cutting-edge technology and transportation infrastructure, revolutionizing how we navigate and manage the movement of people and goods. At its core, ITS involves a sophisticated network of sensors, cameras, communication systems, and advanced algorithms to enhance the efficiency, safety, and sustainability of transportation systems. In the dynamic landscape of such modern transportation, ITS has a direct impact on everyone by optimizing crucial tasks like traffic flow and route optimization while ensuring high safety standards (Sussman (2008)). By leveraging real-time data and intelligent algorithms, ITS facilitates the seamless coordination of vehicles, pedestrians, and other participants in the transportation ecosystem.

The rapid advancements in deep learning technology over recent years have significantly impacted real-time systems like ITS, opening up new possibilities for tasks previously deemed not feasible. Traffic participant detection, an essential task in ITS, aims to provide information assisting counting, speed measurement, identification of traffic incidents, traffic flow prediction, etc. However, to be truly effective, detection methods must meet stringent criteria: they must operate in real-time, withstand variations in lighting and weather conditions, and minimize storage requirements. For instance, in the span of just 1 s, a vehicle may travel over 8.3 m on average, and a pedestrian could cover over 1.43 m, leading to potential misses in fast-paced traffic scenarios and introducing motion blur concerns (Zhang *et al.* (2023)). Moreover, nighttime and different weather conditions make the detection task more challenging as many features such as edge, corner, and shadow do not work due to varying illu-

mination. Detectors tend to rely on features like headlights, rearlights, and beams to detect vehicles. A workaround to detect traffic participants in such cases is to rely on specific detectors for specific cases, eg. a detector that learns texture information for daytime and another detector for utilizing tail-light information for nighttime. However, it is very difficult to develop a universal method for detection in varied lighting conditions (Yang and Pun-Cheng (2018)). Pedestrian detection becomes an even more challenging task since they do not have such illuminating features that the detectors could use (Ghari *et al.* (2024)). In the face of these limitations and the recent algorithmic developments, there exists immense potential in exploring various sensor technologies.

1.1 Motivation and Contribution

The integration of event-based cameras into existing ITS infrastructure holds great promise for robust traffic participant detection in real-time scenarios. Event-based cameras operate on a fundamentally different principle compared to conventional frame-based cameras, capturing asynchronous and continuous streams of pixel-level brightness changes instead of traditional still frames at fixed frequencies. Each "event" is represented by a tuple $\langle x, y, p, t \rangle$ corresponding to an illuminance change by a fixed relative amount at pixel location (x, y) and time t, with the polarity $p \in \{0, 1\}$ indicating whether the illuminance was increasing or decreasing. This approach not only reduces computational throughput burdens and storage requirements but also enhances sensitivity to motion and dynamic events. Event cameras achieve exceptional temporal resolution (over $10K$ fps) and high dynamic range (above 120 dB) (Rebecq *et al.* (2021)), prompting explorations into visual perception and robotics (Gallego *et al.* (2019); Zheng *et al.* (2023)), and its various applications in ITS (Rodríguez-Gomez *et al.* (2020); Tomy *et al.* (2022); Gallego *et al.* (2019)).

Existing multimodal traffic datasets from sensors such as RGB cameras, LiDAR, and Radar have been utilized for several traffic detection tasks in the context of autonomous vehicles (AV) (Geiger *et al.* (2012); Sun *et al.* (2019); Caesar *et al.* (2020); Chen *et al.* (2020)). However, a largely unexplored yet promising domain lies in the use of event cameras for traffic detection for traffic monitoring. This serves as an inspiration to contribute a first-of-its-kind, fully event-based traffic monitoring dataset.

In this work, I present *eTraM* (Verma *et al.* (2024)), a novel, fully event-based traffic perception dataset curated using the state-of-the-art high-resolution Prophesee EVK4 HD event camera (EVK (2023)). The dataset spans over 10 hr of annotated event data, provided from a fixed perspective that facilitates comprehensive traffic monitoring. Experts from the local transportation department were consulted, and an event camera was strategically mounted over selected sites (intersections, roadways, and local streets) to collect traffic data under diverse conditions. The data collection process was systematically conducted across various weather and lighting conditions spanning challenging scenarios such as high glare, overexposure, underexposure, nighttime, twilight, and rainy days. *eTraM* possesses annotations that include over 2M bounding boxes of traffic participants such as vehicles (cars, trucks, buses, trams), pedestrians, and various micro-mobility (bikes, bicycles, wheelchairs) as shown in Figure 1.1. *eTraM* offers the perspective of a static camera captured at a variety of scenes and varying elevations, further enhancing its versatility and applicability in real-world scenarios. This comprehensive approach ensures that *eTraM* captures not only the routine dynamics of traffic but also the nuances and challenges presented by a broad spectrum of scenarios and participants. The richness and breadth of the dataset are tested through various experiments, and its generalization on nighttime and unseen scenes has been evaluated. *eTraM* stands as a valuable re-

source, propelling research and innovation in the evolving field of event-based traffic perception in ITS.

The contributions of this thesis can be summarized as follows:

1. I introduce *eTraM*, a first-of-its-kind fully event-based dataset from a static perspective. The dataset encompasses a diverse variety of traffic scenarios (scenes, weather, and lightning conditions) and participants (vehicles, pedestrians, and micromobility users), with over 2M bounding box annotations for detection and tracking tasks.

2. Establish detection and tracking baselines using state-of-the-art event-based approaches on *eTraM* across the various traffic monitoring scenarios and lighting conditions.

3. Quantitatively evaluate how sensor invariance to absolute illumination affects the generalization capabilities of event-based approaches on nighttime data.

4. Quantitatively analyze the impact of unseen scenarios and variation in object size on event-based data.

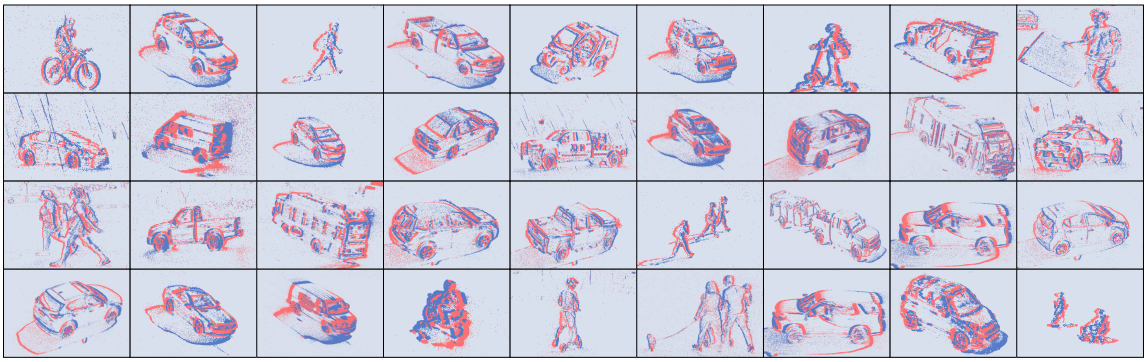

Figure 1.1: Unveiling The Dynamic World Of Road Traffic: A Glimpse Into Our Event-based Traffic Monitoring Dataset Featuring Diverse Traffic Participants, Including Pedestrians, Various Sized Vehicles, And Micro-mobility Users That Include Cyclists, Wheelchair Users, And Bikers.

Chapter 2

BACKGROUND

For many years, traditional frame cameras have been widely used for various tasks. These applications have seen a significant boom over the last few years, thanks to the availability of large amounts of data and computing resources. This has been instrumental in recent breakthroughs, especially in areas that involve natural language processing and generative artificial intelligence (AI). However, for more real-time tasks, they suffer from a bandwidth latency trade-off, which further affects their performance under bad lighting conditions and rapid movements. Frame cameras with high frame rates come at the expense of bandwidth overhead and increased costs, hindering their efficacy in real-world applications.

To address these constraints, researchers have explored alternative sensor technologies such as RGBD, LiDAR, and Radar, seeking to complement the capabilities of traditional frame cameras. Subsequently, numerous studies have discussed the necessity of diverse sensors while deliberating on the significance of various sensor combinations. For instance, the work by Harley *et al.* (2022) examines the boost that radar data can provide to camera-only infrastructure for bird's eye view perception and explores its trade-offs and gaps with the more expensive LiDAR-enabled systems.

However, a promising avenue that remains relatively underexplored is using event cameras in Intelligent Transportation Systems, especially for real-time long-term monitoring. These sensors, which operate on a fundamentally different principle than traditional cameras, offer the potential to overcome the bandwidth-latency trade-off by only capturing changes in the scene, thus reducing the data load and enabling more efficient processing.

2.1 Neuromorphic Sensing

Neuromorphic sensors, commonly also known as event or dynamic vision sensors (DVS), represent a paradigm shift in the field of computer vision. Figure 2.1 (a) is the IMX636HD camera realised through the collaboration between Prophesee and SONY (EVK (2023)). Unlike traditional frame-based cameras, which capture entire scenes at fixed intervals, event cameras are inspired by the efficient and selective processing observed in biological vision systems. Instead of continuously capturing entire scenes, event cameras mimic the behavior of the human retina, detecting and reporting individual pixel-level changes in brightness asynchronously and with remarkable temporal precision.

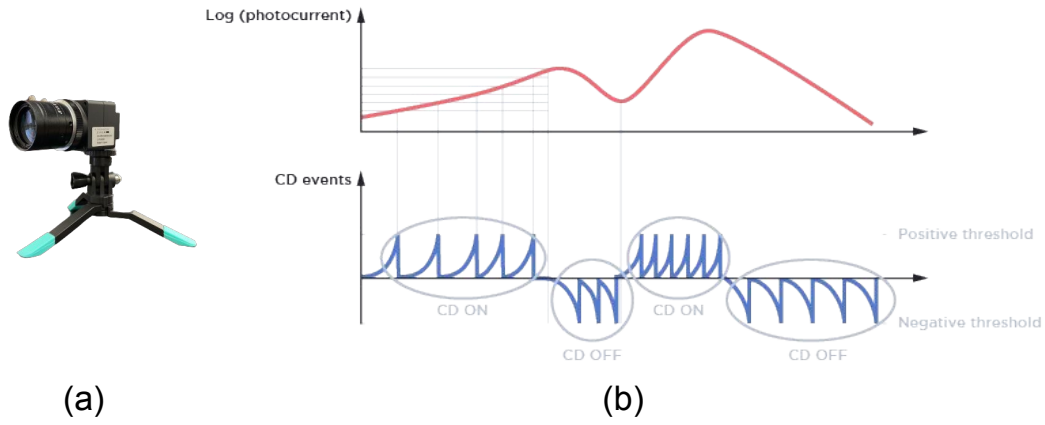

Figure 2.1: Event Generation Model: (a) Is The Compact IMX636HD Event Camera By Prophesee And SONY, And (b) Graphically Illustrates (Prophesee S.A (2024a)) How CD ON (Positive Polarity) And CD OFF (Negative Polarity) Events Are Triggered Due To A Change In The Log Of Photocurrent.

Just as our eyes focus only on relevant changes in our field of view, event cameras are designed to selectively detect significant changes in pixel intensity, allowing event cameras to operate in **real time**. Each pixel at the spatial coordinates $\langle x, y \rangle$ of

the cameras is responsible for triggering an event $\langle x, y, t, p \rangle$ when the logarithmic intensity change L at that pixel exceeds a predefined constant threshold C. The above condition can be mathematically represented as,

$$L(x, y, t) - L(x, y, t - \delta t) = \mathrm{p} \times C \tag{2.1}$$

Here, $t - \delta t$ is the time when the last event at that pixel was triggered, and $p \in \{+1, -1\}$ is the polarity of the event. A positive or negative polarity event is generated based on whether there was an increase or decrease observed at that pixel. The working principle is illustrated in Figure 2.1 (b). Every time the value of $log(photocurrent)$ increases by fixed amounts, consecutive *CD ON* events are triggered. Similarly, *CD OFF* events are triggered for every threshold amount of decrease in the value of $log(photocurrent)$. This threshold can be manually configured depending on the sensor and is often referred to as the *bias*. It can be used to determine the sensor's sensitivity to the change in photocurrent and the rate at which they are allowed to occur.

The output of the event camera, a stream of $\langle x, y, p, t \rangle$ events, can be visualized in a two-channel representation within a three-dimensional space. Here, two dimensions constitute the spatial component capturing the location of the event in the image co-ordinates of the scene, while the third dimension represents the temporal coordinates, indicating precisely when the event occurred. This spatial-temporal representation facilitates efficient processing and analysis of dynamic scenes, enabling tasks such as object tracking, motion estimation, and scene reconstruction with high speed and accuracy.

Unlike traditional cameras constrained by fixed frame rates, event cameras adapt asynchronously to changes in the scene, ensuring that no crucial information is missed. Meanwhile, a frame camera's output consists of a sequence of static frames captured

at regular intervals, leading to information loss between consecutive frames, especially in dynamic scenes. Conversely, the output stream of an event camera showcases a continuous flow of events, preserving every significant change in the scene without any loss of information. This stark contrast in the output streams of event cameras and traditional frame cameras is illustrated in Figure 2.2. The continuous stream of events enables event cameras to excel in scenarios with rapid motion or dynamic lighting conditions. On the other hand, in cases where the object is moving at a high speed, a motion blur is induced in the frames due to the relatively longer exposure time. This is usually tackled by increasing shutter speed, but this comes at the cost of increased redundancy as well as increased bandwidth requirements. This example is illustrated in Figure 2.2 (b), where event cameras continue to smoothly capture such cases due to their working principle. This is made possible due to the **high temporal resolution** of event cameras, with events being detected at resolutions greater than $10K$ fps and asynchronous **pixel latency** lower than 10K microsecond.

Another compelling advantage of event cameras lies in their ability to avoid capturing redundant information and thus have a **low power** requirement. Traditional frame cameras may often waste resources by capturing unchanged regions of the scene, leading to unnecessary data processing and storage. This is highly disadvantageous in long-term monitoring applications. In contrast, event cameras only transmit information when there is a change in pixel intensity, effectively eliminating the need to capture and process unchanged regions of the scene continuously, conserving bandwidth and power. This selective approach not only enhances efficiency but also enables event cameras to perform exceptionally well in scenarios with minimal motion, where traditional cameras may struggle to deliver satisfactory results.

Very importantly, their exceptionally **high dynamic range**, exceeding 120 dB, significantly surpasses the 60 dB range of high-quality, frame-based cameras. This

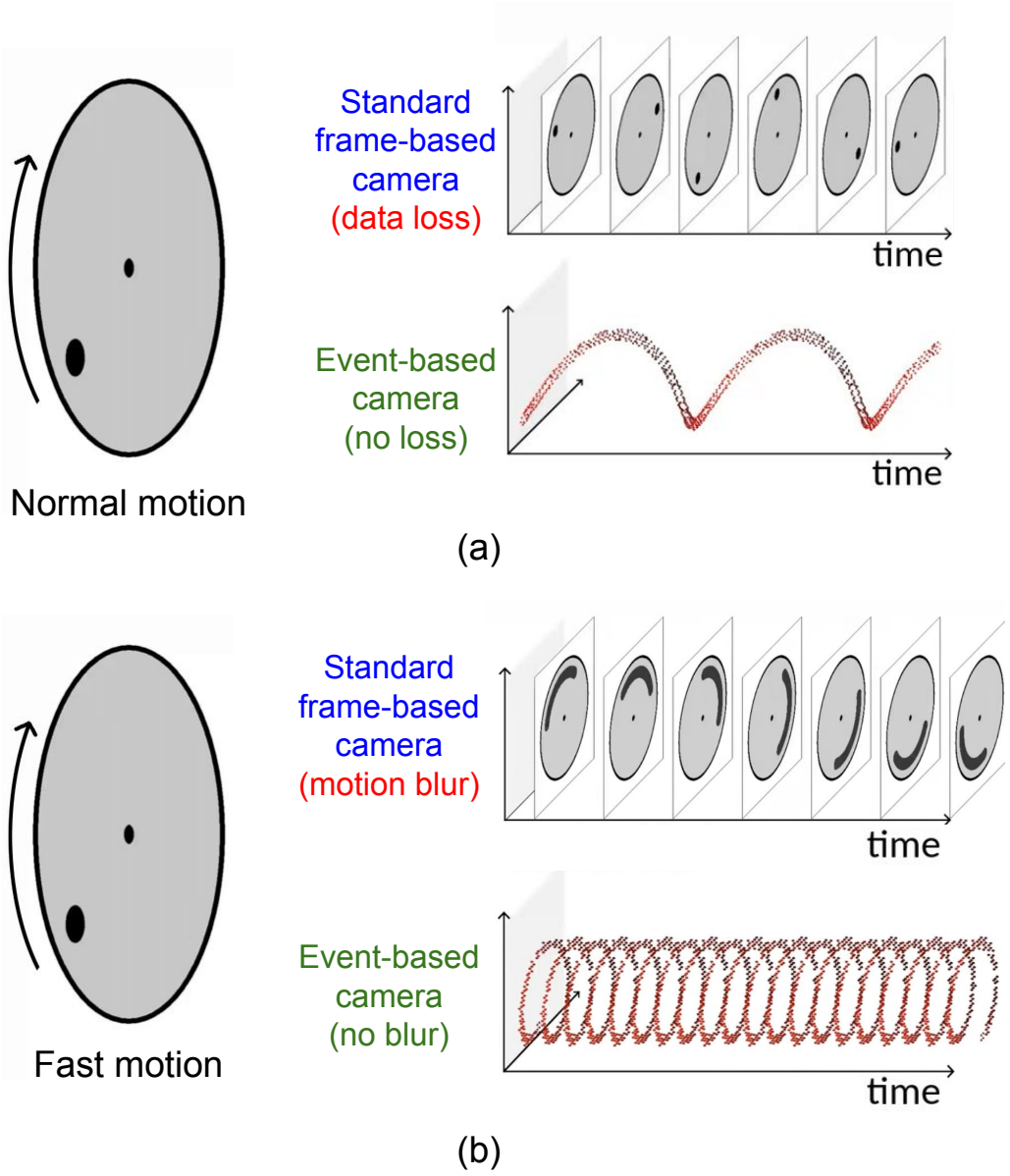

Figure 2.2: Contrasting Standard Camera With Event Camera Illustrating High Temporal Resolution: (a) Demonstrates The Data Loss In Standard Cameras Between Consecutive Frames (redrawn From Mueggler *et al.* (2014)). Meanwhile, (b) Illustrates Motion Blur Typical In Frame-based Cameras During Fast Motion, A Phenomenon Absent In Event-based Cameras.

enables them to capture information across a broad spectrum, from moonlight to daylight. This superior range is attributed to the logarithmic scale operation of the photoreceptors in the pixels and the independent functioning of each pixel, which eliminates the need for a global shutter. Similar to biological retinas, event camera pixels exhibit adaptability to both very dim and extremely bright stimuli.

2.2 Event-based Datasets

Despite the tremendous potential of event cameras, the scarce availability of datasets has been one of the biggest hurdles in the advancement of event-based vision for various applications. However, with the increasing accessibility of event-based sensors, a notable rise in datasets and, consequently, in research and development of event-based vision has been seen. This section aims to provide an overview of the evolution of event-based datasets over time, with a specific focus on datasets relevant to Intelligent Transportation Systems (ITS).

2.2.1 Early event-based datasets

Due to the lack of commercially available event cameras and high costs, **early event-based datasets** often involved the transformation of frame-based datasets into event streams. A noteworthy example is the work in Orchard *et al.* (2015), where MNIST (Lecun *et al.* (1998)) and Caltech-101 (Fei-Fei *et al.* (2006)) datasets were converted to event streams by moving an event camera in front of a screen displaying frame data. These datasets proved themselves useful for benchmarking various event-based algorithms; however, converting frame-based datasets still required event cameras, which were not easily available. Later works proposed event simulators like ESIM (Rebecq *et al.* (2018)), vid2e (Gehrig *et al.* (2020)) and v2e (Hu *et al.* (2021)). The advantage of these simulators was their ability to leverage any existing

widely-used frame-based dataset and model them into their event-based counterparts without the need to own an event camera. There also exists a DVS camera sensor in the CARLA simulator (Dosovitskiy *et al.* (2017)) that allows the simulation of specific situations and scenarios in a controlled manner.

2.2.2 Ego-motion event-based datasets

Due to their temporal prowess, research in event-based vision has attracted the interest of many researchers from an ego perspective where quick response is key. Hence, with an increase in the accessibility of event cameras in recent years **ego-motion event-based datasets** have seen a rise. The initial efforts in deploying event cameras for driving scenarios was pioneered in the works in DDD17 Binas *et al.* (2017), and DDD20 Hu *et al.* (2020) using a 346x260 pixels DAVIS sensor. These datasets focused on steering angle prediction, with DDD17 having 12 hours of driving data. Later, DDD20 extended DDD17 to have a total of 51 hours of driving data. MVSEC Zhu *et al.* (2018a) presents a multimodal stereo dataset fusing 346x260 DAVIS sensors along with LiDARs, IMUs, and RGB cameras for three-dimensional perception tasks, such as feature tracking, visual odometry, and stereo depth estimation, marking the first work to involve event-cameras from a multi-sensor fusion approach. DSEC Gehrig *et al.* (2021) further expands these fusion efforts by including 390K annotations for detection tasks on an hour of multimodal stereo data using 640x480 pixels Prophesee Gen3.1 sensors.

Prophesee Prophesee (2023) introduced two substantial ego-motion datasets for detection tasks in quick succession, the Gen1 Automotive dataset (de Tournemire *et al.* (2020)) and the 1 Megapixel Automotive dataset (Perot *et al.* (2020)) using their own manufactured cameras. The Gen1 Automotive Detection Dataset de Tournemire *et al.* (2020) encompasses a total of 255,781 manually annotated bounding boxes

(228,123 cars and 27,658 pedestrians instances) acquired over a span of 39 hours using 304×240 pixels Prophesee Gen1 sensor. 1 Megapixel Automotive Dataset Perot *et al.* (2020) stands out as the most comprehensive ego-motion event-based detection dataset. It encompasses 15 hours of recorded footage, featuring a resolution of 1280×720 pixels, with 25 million generated bounding boxes. However, they are unable to provide extensive nighttime annotations due to their automated labeling protocol. In 2023, the PEDRo dataset (Boretti *et al.* (2023)) was released, which was the first event-based dataset that focuses on people detection from a robotics perspective. It contains 43,259 bounding boxes from 119 recordings with an average duration of 18s.

2.2.3 Fixed perception event-based datasets

Ego-motion datasets capture events from the background as well due to their relative motion with respect to the moving camera, thus leading to datasets from an ego-motion perspective sacrificing the sparse nature of event-based data. Despite this not occurring when cameras are not moving, **fixed perception event-based datasets** have been few. Datasets such as DVS-Pedestrian (Miao *et al.* (2019)) are limited to pedestrian detection using a static 346x260 pixels DAVIS346 camera. The dataset has 12 recorded sequences containing 4670 labeled instances of pedestrians. DVS-OUTLAB (Bolten *et al.* (2021)) explores the plausibility of using event cameras for long-time monitoring purposes. It consists of recordings from three fixed 768x640 pixels CeleX-4 DVS event cameras featuring outdoor urban public areas involving persons, dogs, bicycles, sportsball as objects of interest. While DVS-OUTLAB presents a dataset from an outdoor environment, it does not primarily focus on environments with interactions between various traffic participants.

Dataset Name	Year	Duration	Perspective		Traffic Participants			Lighting			Weather		No. of Bbox	Scenarios
			Ego	Static	VH	PED	MM	Day	Night	Twilight	Clear	Rainy		
DDD17 Binas *et al.* (2017)	2017	12	✓		✓			✓	✓	✓	✓	✓	-	Driving
MVSEC Zhu *et al.* (2018a)	2018	-			✓			✓	✓		✓		-	Driving, Handheld
DVS Pedestrian Miao *et al.* (2019)	2019	0.1		✓		✓		✓			✓	✓	4.6K	Walking street
DDD20 Hu *et al.* (2020)	2020	51	✓		✓			✓	✓	✓	✓		-	Driving
Gen1 de Tournemire *et al.* (2020)	2020	39	✓		✓	✓		✓			✓		255K	Driving
1 Megapixel Perot *et al.* (2020)	2020	15	✓		✓	✓	✓	✓	✓		✓		25M	Driving
DSEC Gehrig *et al.* (2021)	2021	1	✓		✓	✓	✓	✓	✓		✓		390K	Driving
DVS-OUTLAB Bolten *et al.* (2021)	2021	7		✓			✓	✓			✓	✓	47K	Playground
PEDRo Boretti *et al.* (2023)	2023	0.5	✓			✓		✓	✓	✓	✓	✓	43K	Robotics
***eTraM* (Ours)**	2024	10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2M	Intersections, Roadways, Local streets

Table 2.1: A Comprehensive Review Of Event-based Traffic Datasets From 2017 To 2024. (VH - Vehicle, PED - Pedestrian, MM - Micro-mobility)

Table 2.1 is a comprehensive review of event-based traffic datasets providing deeper insights into the duration, perspective, traffic participants, lighting conditions, weather, and size.

2.3 Event Representations

Due to the unique characteristics of event data, a myriad of ideas have been proposed to represent it, depending on the downstream tasks. This section provides an overview of the various ways event data can be processed and the trade-offs incurred in each of them.

2.3.1 Individual Events

Methods for event-by-event processing, such as probabilistic filters and Spiking Neural Networks (SNNs), are employed when dealing directly with events. These filters or SNNs incorporate supplementary information accumulated from past events or provided by external knowledge. This information is asynchronously integrated with

incoming events to generate an output (Yao *et al.* (2021), Zhang *et al.* (2022)). SNNs propagate information sparsely within the network and share similarities with dense recurrent neural networks (RNNs) in that each spiking neuron maintains an internal state that updates over time. Unlike RNNs, however, neurons in SNNs emit spikes only when a certain threshold is exceeded. This non-differentiable spike generation mechanism poses significant challenges in optimizing these networks. One approach to mitigate this issue is to circumvent the threshold and instead propagate features across the receptive field (Messikommer *et al.* (2020)). However, this modification sacrifices the sparse-processing property in deeper layers of the network. Consequently, the design and training complexity of SNNs necessitates further exploration and investigation before achieving competitive performance.

2.3.2 Image/Tensor Representation

Due to the extensive research on frame-based data, it is advantageous to convert event data into a representation that can be leveraged by these architectures. An important way to achieve this representation is by using the histogram of events. It involves assigning each event to a specific cell based on its position (x, y) and a time bin determined by its timestamp (t).

In Maqueda *et al.* (2018), the authors define the histogram as the total count of events that occurred in the corresponding spatial cell within each time bin. However, the count is done separately for each polarity, which results in a total of two output channels. Let H represent a four-dimensional tensor with dimensions n, c, h, w, where n represents the index of the timestamp, c represents the channel for the two polarities, h represents the height, and w represents the width of the input event stream. Every new event $\langle x, y, p, t \rangle$ corresponds to a specific histogram decided by the time bin that the timestamp corresponds to. Next, the histogram is updated by adding 1 at the

spatial coordinates of the new event. The mathematical representation of the update is as shown in Equation 2.2,

$$H(\tfrac{t}{\Delta}, p, y, x) = H(\tfrac{t}{\Delta}, p, y, x) + 1, \tag{2.2}$$

where Δ is the time interval.

However, this method leads to the representation losing the fine-grained temporal resolution in the event data. To retain some temporal information in each tensor, several methods consider the contribution of a new event as a function proportional to the proximity of the time at which the new event occurred with the time bin corresponding to the tensor.

Since there can also be a sudden and uneven trigger of events within the fixed time intervals, it is possible for some bins to be very dense while some to be extremely sparse. To tackle such cases, Wang *et al.* (2019) propose to stack events by a strategy that considers a fixed number of events rather than a fixed time interval, as discussed previously. Some works (Liu and Delbrück (2018); Liu and Delbruck (2022)) have explored the efficacy of dynamic adjustment of exposure time and inter-slice time intervals. They show that the adaptable control mechanism enhances model robustness in dynamic environments characterized by diverse motion speeds and scene structures.

2.3.3 Time-Surfaces

The time surface, an alternative event processing method, involves recording the timestamp of the most recently received event for each pixel. This technique considers polarities independently, resulting in the output of two channels (Lagorce *et al.* (2017)). By doing so, the representation considers the rich temporal information of the events and can be updated asynchronously.

However, since this gives equal importance to older events, some works incorporate

an exponential decay to the timestamps to diminish the influence of older events. Assuming $t_0 = 0$ for simplicity, this decay process can be easily implemented. The input representation is represented as a three-dimensional tensor $\langle p, w, h \rangle$, where p represents the polarity, h represents the height, and w represents the width of the input event stream.

For each event $\langle x, y, p, t \rangle$ when $t \leq t_i$, its contribution to the time surface at time t_i can be mathematically represented as shown in Equation 2.3,

$$TS_{t_i}(p, y, x) = exp(-\frac{t_i - t}{\tau}), \tag{2.3}$$

where τ is the normalization constant (Sironi *et al.* (2018)). As time surfaces only store a single value corresponding to the latest event, they compress information well. However, this also leads to a degraded representation in scenes with a lot of textures since the pixels spike frequently.

2.3.4 Voxel-based

Voxel-based representations translate raw events into the nearest temporal grid within temporal bins. The concept of the first spatial-temporal voxel grid was introduced in Zhu *et al.* (2018b), which involves inserting events into volumes using linearly weighted accumulation to enhance temporal domain resolution. Voxel-based representations are often utilized in 3D object recognition, reconstruction, and scene understanding tasks, especially when dealing with point clouds obtained from depth sensors like LiDAR or RGB-D cameras. These representations are valuable for tasks where spatial relationships and volumetric information are crucial, such as autonomous driving, robotics, and augmented reality applications. Lately, there has been an increased interest in leveraging these representations for event data with Baldwin *et al.* (2022) proposing a time-ordered recent event (TORE) method aiming to maintain spike

temporal information with minimal information loss.

2.3.5 Graph-based

A recent line of work tries to exploit the spatio-temporal characteristics of events by modeling the raw event data as a graph for various downstream tasks like object classification (Mesquida *et al.* (2023)), detection (Schaefer *et al.* (2022)), and optical flow estimation (Dalgaty *et al.* (2023)). The graph structure is created based on the position of the events in the 3-dimensional coordinate frame. Each event may act as a node in the graph, and edges may be formed based on its proximity to other nodes. The approach aims to exploit the sparsity of raw event data by transforming events within a time window into a set of connected nodes. However, a major challenge is to design architectures in a way that information can propagate over vast distances in the time dimension. This is especially important when large objects move slowly with respect to the camera or momentarily halt.

Chapter 3

THE *ETRAM* DATASET

eTraM, short for Event-based Traffic Monitoring, is a dataset with applicability in event-based long-term and resource-efficient traffic monitoring from a static perspective. This chapter outlines *eTraM*'s acquisition framework, preprocessing techniques, annotation strategies, and statistics, providing deeper insights into the dataset and its annotations.

3.1 Dataset Acquisition Framework

To capture high-quality data, the Prophesee EVK4 HD event camera (EVK (2023)), notable for its high resolution (1280×720 px), high temporal resolution (over $10,000$ fps), dynamic range (above 120 dB), and exceptional low light cutoff (0.08 Lux) has been used. The sensor was strategically positioned at a height of approximately 6 m with a pitch angle of about 35° to the ground. The configuration is deliberately chosen this way to maintain consistency with the placement of traffic cameras in existing infrastructure and to ensure comprehensive coverage of interactions between diverse traffic participants.

The dataset comprises recorded sequences obtained at multiple intersections, roadways, and local streets around Arizona State University, Tempe Campus. The data sequences were recorded for intervals of $15 - 30$ min at different times of the day, covering daytime, nighttime, and twilight. The dataset also observes different weather conditions, including sunny, overcast, and rainy. To achieve this, extensive data collection efforts were carried out over a span of 8 months. Figure 3.1 shows different data acquisition sites considered for data collection.

Figure 3.1: Data Collection Setup: The First Four Images From The Top Left Display Daytime Data Collection Sites, The Center Image Shows The Prophesee EVK4 HD Camera And The Last Four Images Depict Nighttime Collection Sites.

3.2 Preprocessing and Annotation

Given the sensitive nature of the event sensor, it was observed that nighttime data tends to exhibit higher levels of noise, primarily attributed to reflections and pointed sources of light from streets and vehicles. To address this challenge and enhance the quality of the data, the recorded sequences are passed through a spatiotemporal filter (Brosch *et al.* (2015)). This spatiotemporal filter works on the idea that events from real objects should occur closer together in both space and time more often compared to events from random noises (Gallego *et al.* (2019)).

The filtering mechanism discards an event if a threshold amount of events with the same polarity do not occur within a fixed temporal window in the vicinity of its 8-neighborhood spatial coordinates. For any $e = \langle x, y, p, t \rangle$, this condition is represented mathematically in Equation 3.1

$$\sum_{t_i=t}^{t+\partial t} \sum_{x_i=x-1}^{x+1} \sum_{y_i=y-1}^{y+1} P(e_i = \langle x_i, y_i, p_i, t_i \rangle, e) > n_{thres}, \tag{3.1}$$

where ∂t represents the temporal window and P(e_i, e) equals 1 only when the polarity e_i and e are the identical. A sample visualization demonstrating the significance of spatiotemporal filtering is illustrated in Figure 3.2.

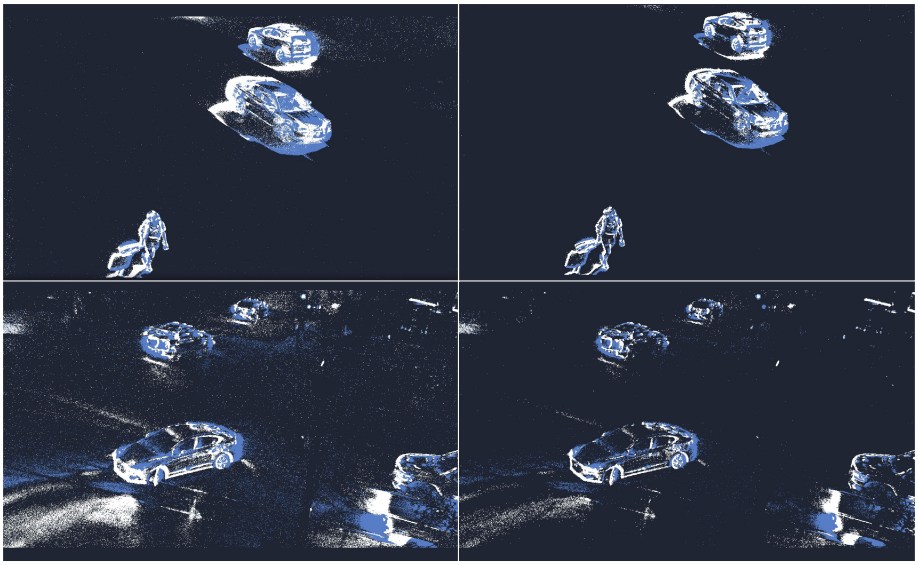

Figure 3.2: Impact Of Spatiotemporal Filtering On Event Camera Data: Comparison Of A Noisy Pre-filtered Image (Left) And The Enhanced Clarity Achieved Post-filtering (Right) On Daytime (Top Row) And Nighttime Data (Bottom Row)

For experiments in future chapters, a temporal window of 10 ms with a minimum threshold of 2 neighboring events has been chosen. The specific filter values were determined through comprehensive experiments detailed in Bolten *et al.* (2021) and further validated through a smaller experiment conducted during the training phase.

Following the denoising stage, events within the stream are partitioned into discrete time bins and consolidated into a single frame, thereby converting the asynchronous event stream into synchronous frames of 30 Hz. These frames are then annotated using CVAT (Sekachev *et al.* (2020)), an open-source annotation tool. The rigorous manual annotation process resulted in the precise identification of 2 M 2D bounding boxes. The format of each event and bounding box annotation is explained in Table 3.1 and Table 3.2, respectively.

Key	**Description**
x	x coordinate of the event in image coordinate frame
y	y coordinate of the event in image coordinate frame
p	polarity corresponding to event $(+1/-1)$
t	time stamp of the event occurrence (µs)

Table 3.1: Format Of An Asynchronous Event.

ID	**Key**	**Description**
0	t	*Timestamp* of the bounding box
1	x	x coordinate of the top left point of the bounding box
2	y	y coordinate of the top left point of the bounding box
3	w	*width* of the bounding box
4	h	*height* of the bounding box
5	class_id	*Class ID* of the object
6	track_id	*Tracking ID* of the object
7	class_confidence	*Confidence score* of detection

Table 3.2: 2D Bounding Box Annotation Format In *ETraM*.

3.3 Dataset Statistics

Here, the key characteristics of the collected data and annotations are highlighted. The dataset encompasses three distinct traffic monitoring scenarios with 5 hr of intersections, 3 hr of roadways, and 2 hr of local streets. Each scenario is collected at multiple locations. For instance, the intersection scenario contains data from 2 four-way, a three-way, and an uncontrolled intersection. Each location has daytime, twilight, and nighttime data totaling 10 hr of data with 5 hr of daytime and 5 hr of nighttime data. The dataset contains 2 million instances of 2D bounding box annotations for traffic participant detection tasks. These annotations additionally include object IDs, making it possible to track objects. The annotation classes encompass a range of traffic participants, from pedestrians and various vehicles (cars, trucks, buses, and trams) to the inclusion of micro-mobility (cyclists, wheelchair users, and bikers).

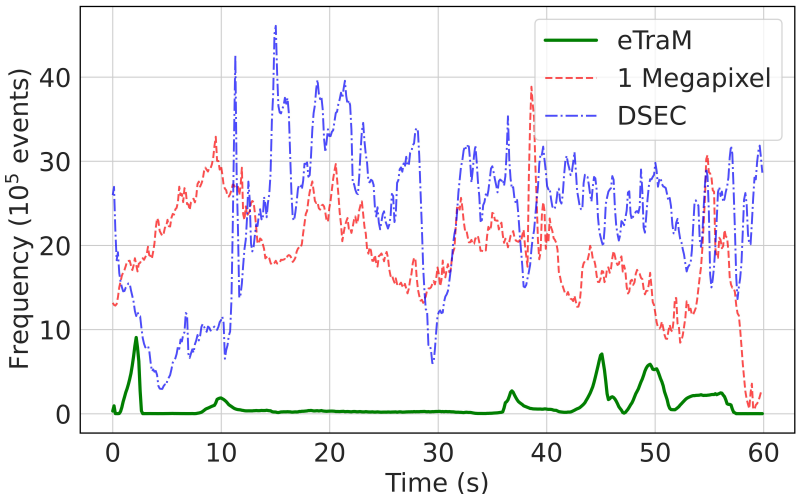

Figure 3.3: A Histogram Illustrating The Event-time Frequency Of *ETraM* (Static Event Dataset) As Compared To 1 Megapixel And DSEC (Ego-motion Event Datasets).

A comparison of the event distribution in *eTraM* with other ego-motion event-based traffic datasets like 1 Megapixel Automotive and DSEC for 60 sec is performed to gain insights into the inherent properties of the dataset. It was observed that the number of events in *eTraM* was significantly lesser by a factor of 30, as shown in Figure 3.3. This is accredited to the static nature of the camera in *eTraM*, which primarily focuses on moving traffic participants in a scene. In contrast, other datasets from an ego-motion perspective capture more data due to the relative motion of the surrounding infrastructure, leading to a continuous and dense stream of events. This sparsity of events in *eTraM* data, combined with the asynchronous nature of events, leads to low memory utilization. This is particularly advantageous for the memory-limited devices used in traffic monitoring infrastructure.

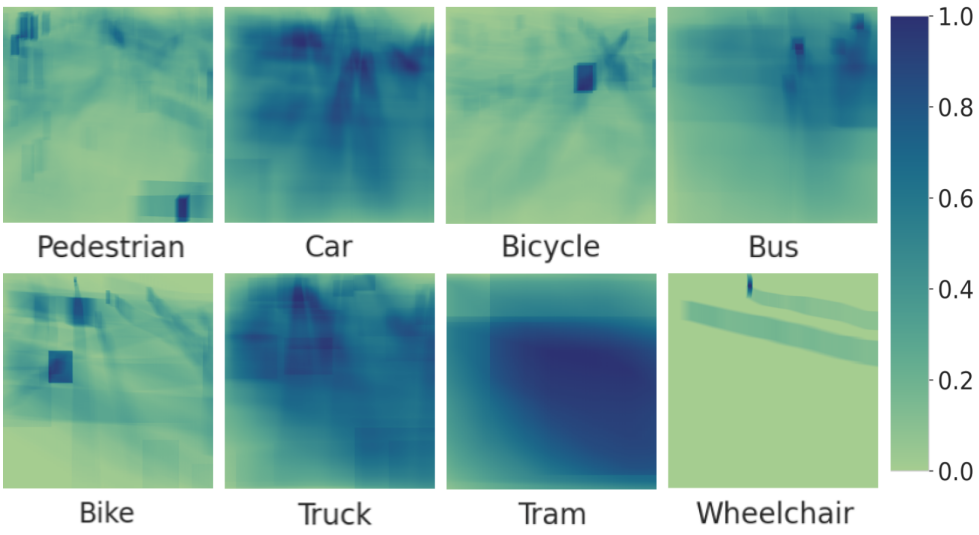

Figure 3.4: The Object Density Of Various Classes Across The Frame.

Figure 3.4 illustrates the spatial distribution of each class within the frame. A uniform spread of vehicle classes across the entire frame and the pedestrians class

covering more than 50% of the frame is observed. This is unlike that observed in various ego-motion event-based datasets, which predominantly feature pedestrians on the two corners of the frame, which often correspond to the sidewalk. This varied distribution also safeguards the model from developing a bias for certain classes in a specific region of the frame.

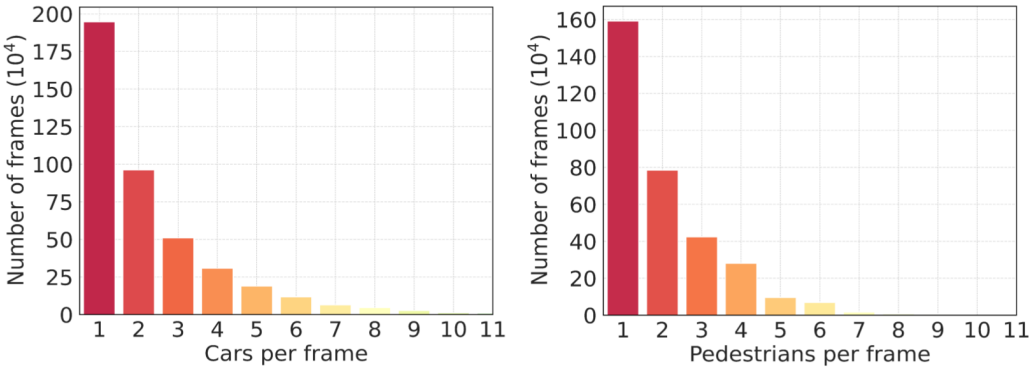

Figure 3.5: Power-law Distribution Of The Number Of Instances Within An Image For Most Predominant Classes - Cars And Pedestrians.

To understand more about the occurrences of the instances per frame, the distribution of the major classes - cars and pedestrians is illustrated in Figure 3.5. While both classes demonstrate a power-law distribution with an increase in the number of instances per frame, the higher number of cars consistently observed per frame indicates that *eTraM* has a dense presence of vehicular traffic compared to pedestrian activity. This dense presence suggests a higher reliance on vehicular transportation within the various locations. These observations could be due to the locations chosen in *eTraM*. Such insights into the distribution of instances per frame can be valuable for understanding traffic patterns, urban planning, and resource allocation for transportation infrastructure and safety measures.

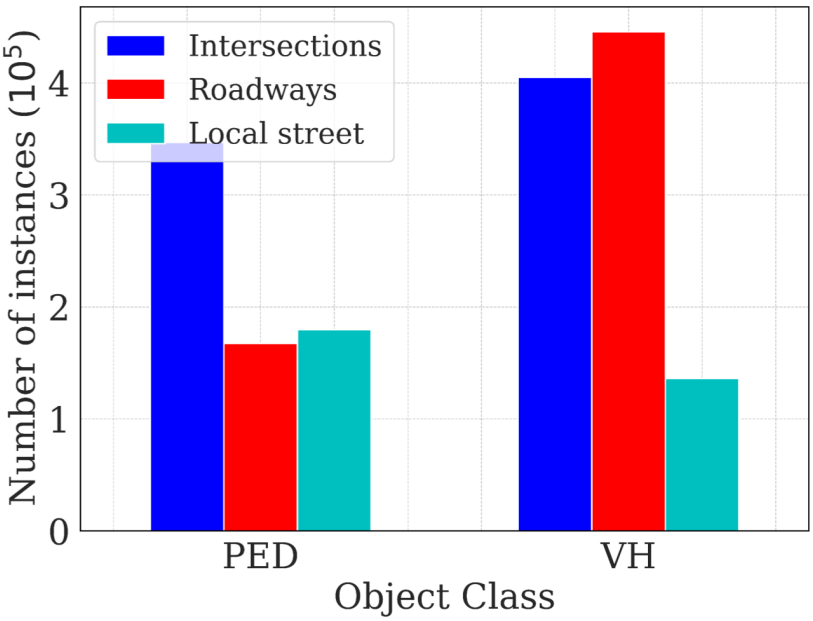

Figure 3.6: Distribution Of Two Major Traffic Participant Categories Across Various Traffic Sites.

While providing an overview of the overall presence of pedestrian and vehicle classes in the dataset, these observations prompt a deeper exploration into variations of the classes across different traffic locations in the dataset. Figure 3.6 illustrates the scene-wise distribution of categories at different traffic locations. This also provides a representation of the real-world dynamics of the major traffic participants at these locations. Pedestrians appear to be mostly populated at intersections but witness a substantial drop in numbers at roadways. Conversely, there is a slight increase in the number of vehicles observed on roadways, showcasing the difference between intersections and roadways. In contrast, local streets feature a lesser number of instances from both categories. This also illustrates the importance of monitoring intersections. At such locations, an equal presence of vehicles and vulnerable road users, such as pedestrians, results in significantly higher interactions between various traffic participants, emphasizing the need for monitoring and improved safety measures.

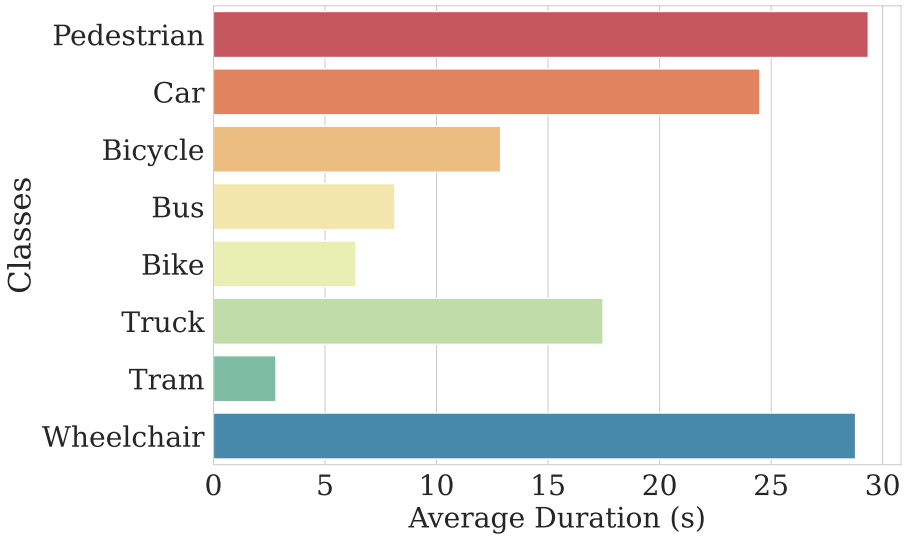

Figure 3.7: The Bar Plot Illustrates The Average Duration, In Seconds, Spent By Instances Of Different Classes, Providing Insights Into The Temporal Characteristics Of Each Class In The Dataset.

Figure 3.7 presents the average duration spent by objects from each class at the traffic site. This temporal analysis sheds light on the distinctive time dynamics of different classes within the dataset. Participants from the pedestrian and wheelchair classes spend the most time at the traffic site, which also corresponds to the speed at which they move. On the other hand, other classes from the vehicle category, including two-wheelers, tend to comparatively spend less time.

Further analysis is performed on the distribution of different categories (VH, PED, AND MM) by the area they cover - small, medium, or large, as shown in Figure 3.8. Based on this grouping, further analysis is conducted in future chapters to get insights into the variation of performance of the event-based detectors with the size of objects.

For accessibility and ease of use, *eTraM* is provided in multiple formats: RAW, DAT, and H5 (Prophesee S.A (2024b)). Additionally, the annotations are available in numpy format. The dataset is split into 70% training, 15% validation, and 15% test-

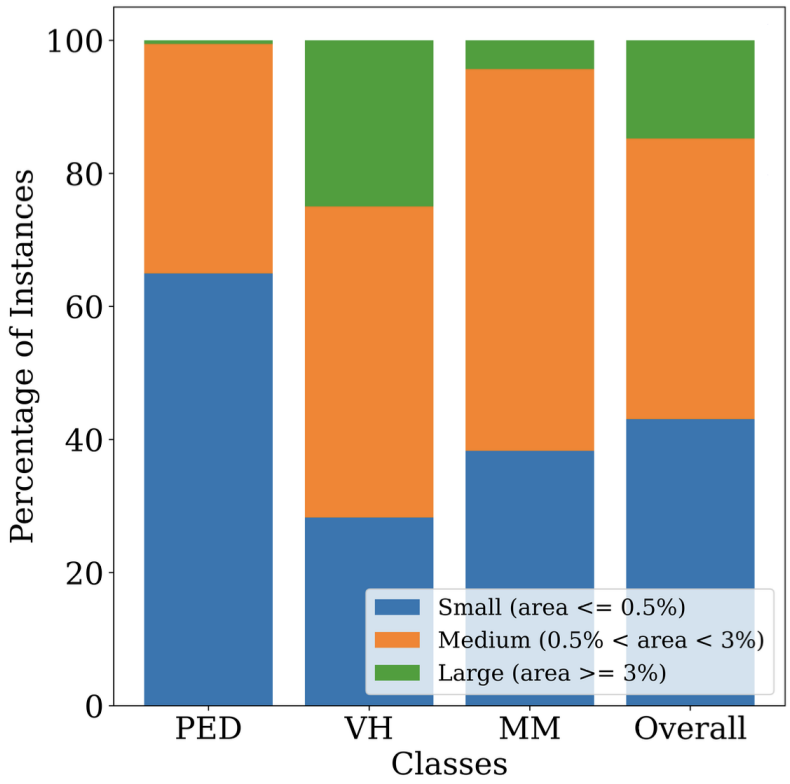

Figure 3.8: The Bar Plot Illustrates The Average Duration, In Seconds, Spent By Instances Of Different Classes, Providing Insights Into The Temporal Characteristics Of Each Class In The Dataset.

ing, ensuring that each subset has proportional data from each scenario. To the best of my knowledge, this stands as a first-of-its-kind event-based dataset from traffic monitoring. To the best of my knowledge, this stands as a first-of-its-kind event-based dataset for traffic monitoring. Additionally, the dataset encompasses nighttime data, enhancing its versatility for a broader range of research applications, some of which are explored in a future chapter.

Chapter 4

BENCHMARKING *ETRAM*

The chapter provides an overview of the baselines established on *eTraM*. The experiments are conducted to evaluate the efficacy of the dataset across diverse scenes and lighting conditions. Two primary tasks, detection (detailed in Section 4.1) and tracking (discussed in Section 4.2), make up the evaluation.

For evaluating *eTraM* on detection tasks, two SoTA tensor-based methods, specifically Recurrent Vision Transformers (RVT) (Gehrig and Scaramuzza (2023)), Recurrent Event-camera Detector (RED) (Perot *et al.* (2020)), and an approach frequently used for RGB frame-based detection tasks, You Only Look Once (YOLOv8) (Jocher *et al.* (2023)) are used. This comparison helps compare how SoTA approaches, utilizing dense tensor representation, perform in comparison to the conventional frame-based approach, which does not use temporal bins. The evaluations are performed on the three major categories of traffic participants - vehicles (VH), pedestrians (PED), and micro-mobility (MM). The mean Average Precision at a 50% Intersection over Union threshold (AP50) is reported for object detection, providing a standardized measure of detection accuracy. For tracking, both Multi-Object Tracking Accuracy (MOTA) and Multi-Object Tracking Precision (MOTP) are reported, offering a holistic assessment of tracking performance in terms of accuracy and precision.

The meticulous evaluation methodology employed in this study ensures robust insights into the performance of *eTraM* across various real-world traffic scenarios, laying a solid foundation for further advancements in event-based traffic monitoring.

4.1 Traffic Participant Detection on *eTraM*

To assess the performance of event-based detectors, the models are trained on 7 hr of data and evaluated on 1.5 hr of validation and test data. RVT and RED were trained from scratch over 3 days on NVIDIA A100, while YOLOv8 was trained for 2 days. Separate evaluations were conducted to provide insights into how each model performs in different scenes and lighting conditions. These values facilitate the understanding of how these models handle diverse and changing contexts where camera placement and environment are drastically different.

Traffic Site	Lighting	RVT				RED				YOLO			
		PED	VH	MM	All	PED	VH	MM	All	PED	VH	MM	All
Intersections	Daytime	**0.460**	0.813	0.315	**0.722**	**0.395**	0.593	0.284	0.545	0.167	0.293	0.111	0.190
Roadways		0.430	0.733	0.070	0.627	0.347	0.590	0.055	**0.551**	**0.173**	0.290	0.004	0.156
Local Streets		0.196	**0.938**	**0.586**	0.316	0.208	**0.875**	**0.695**	0.351	0.124	**0.559**	**0.204**	**0.296**
All Scenes		0.304	0.781	0.403	0.572	0.302	0.656	0.251	0.497	0.142	0.309	0.112	0.188
Intersections	Nighttime	0.161	0.465	-	0.262	0.149	0.425	-	0.242	0.071	0.375	-	0.149
Roadways		0.310	0.827	-	0.739	0.362	0.782	-	0.726	0.004	0.229	-	0.117
Local Streets		**0.739**	**0.868**	0.097	**0.829**	**0.722**	**0.831**	0.145	**0.817**	**0.198**	**0.486**	0.030	**0.239**
All Scenes		0.317	0.674	0.064	0.523	0.303	0.660	0.083	0.504	0.123	0.322	0.013	0.153
Overall		0.309	0.717	0.313	0.539	0.303	0.649	0.197	0.491	0.134	0.314	0.086	0.178

Table 4.1: Baseline Evaluation: Comprehensive Evaluation Of State-of-the-art Tensor-based Approaches RVT, RED, And Frame-based Approach YOLOv8 Across Various Traffic Sites (Intersections, Roadways, Local Streets) During Both Daytime And Nighttime For PED - Pedestrian, VH - Vehicle, And MM - Micro-mobility.

Several key observations that emerged from the evaluation results of tensor-based models are shown in Table 4.1. Notably, the detection of vehicles consistently outperforms pedestrian detection across all scenes and models. The performance on micro-

mobility shows significant variance, likely due to the smaller number of instances and the broad range of subjects captured under the category. During daytime, both vehicle and pedestrian detection results are consistent in intersections and roadways. However, in local streets, vehicle detection improves as compared to other scenes due to fewer instances and reduced occlusion. A decline in pedestrian detection is observed, perhaps due to occlusions caused by high pedestrian densities. During nighttime, vehicle detection in local streets and roadways remains consistent. Interestingly, pedestrian detection improves significantly on local streets at night. This can be attributed to a combination of less pedestrian-to-pedestrian occlusion, instances being closer to the camera, and additional visual features like shadows that become more prominent at night. Due to noise from various light sources, nighttime intersections observe a substantial drop in performance. Despite this, the performance during nighttime is at par with daytime scenarios, with an increase in vehicle detection on roadways at night. This increase could be due to reduced vehicle-to-vehicle occlusion as fewer vehicles are observed in nighttime conditions.

It is observed that YOLOv8 performs poorly compared to RED and RVT models, demonstrating the advantage of using tensor-based representation over conventional frame-based methods. This could be attributed to YOLOv8 not using any temporal information from the event stream. While RVT and RED do not fully exploit the temporal component of event streams, they make use of temporal bins in their preprocessed representations and have recurrent networks constituting their architecture as well. Figure 4.2 and Figure 4.1 show a qualitative example of the detection task on *eTraM*. In summary, the evaluation showcases the relative difficulties of various traffic monitoring scenarios and classes and the strength of event-based detectors in nighttime conditions.

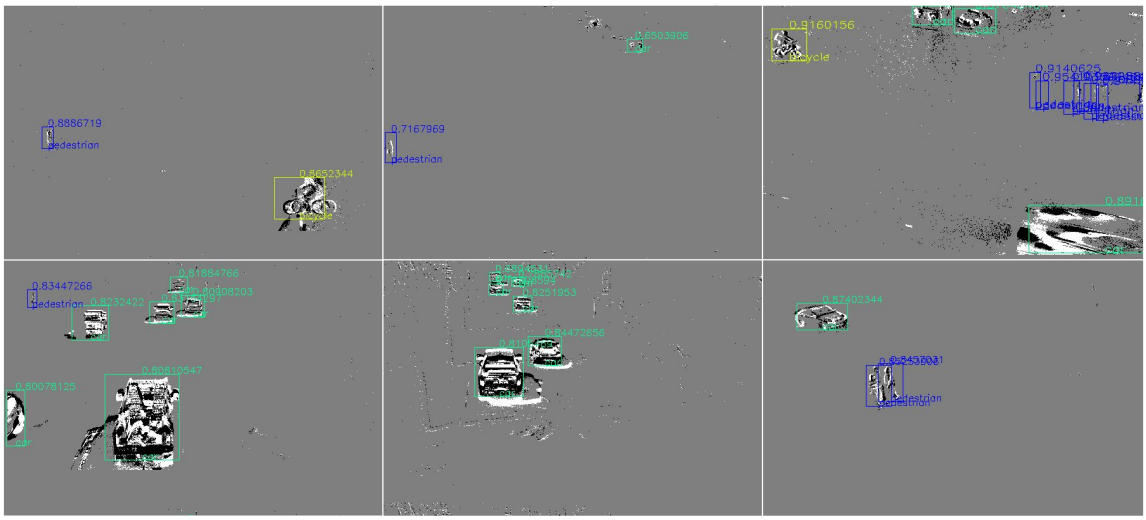

Figure 4.1: Traffic Participant Object Detection By RVT. Snapshots Illustrating The Detection Results Of RVT At Various Traffic Sites, Showcasing Its Performance In Diverse Real-world Scenarios.

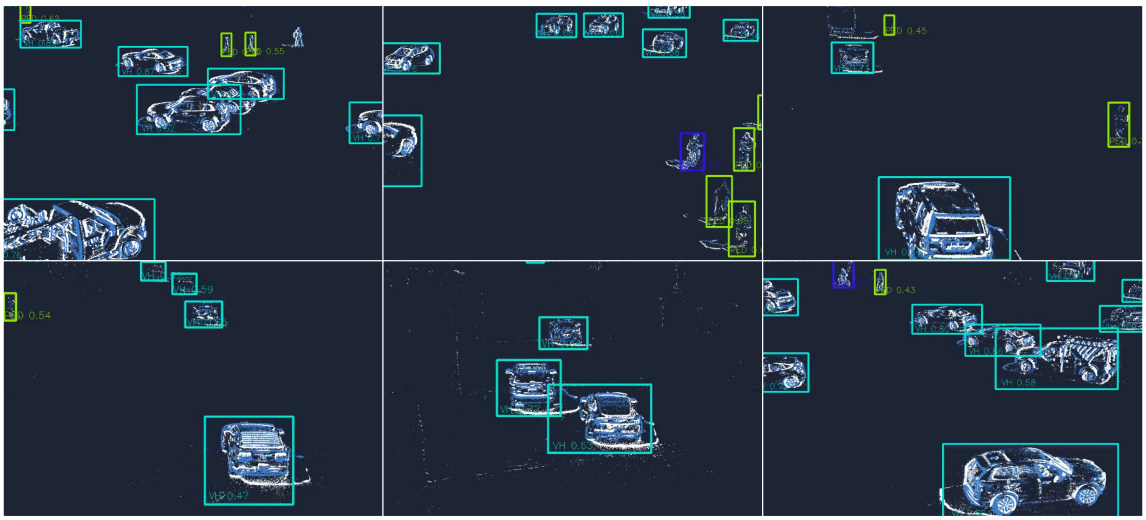

Figure 4.2: Traffic Participant Object Detection By RED. Snapshots Illustrate The Detection Results Of RED At Various Traffic Sites, Showcasing Its Performance In Diverse Real-world Scenarios.

4.2 Multi-Object Tracking on *eTraM*

Tracking plays an important role in a fixed perspective, long-term monitoring scenario. Tracking enables the analysis of object behavior over time. By following the trajectories of objects within the scene, patterns, anomalies, and trends can be identified. This is particularly valuable in scenarios where abnormal behaviors or events need to be detected and investigated. In environments with multiple people, tracking enables the simultaneous monitoring and tracking of multiple individuals. This capability is essential for Re-Identification (ReID) systems deployed in crowded or busy spaces, where tracking and re-identifying individuals amidst a crowd pose significant challenges. Tracking IDs can be crucial for evaluating the tracking ability of a model, particularly in scenarios where multiple objects are being tracked over time. Tracking IDs help associate predictions across different frames, allowing you to measure how consistently and accurately the model maintains the identity of each tracked object. Although evaluation of the tracking abilities of models is not possible in most event-based datasets, *eTraM* enables it by providing ground truth tracking IDs for each object.

The Intersection-over-Union (IoU) based thresholding technique (Bochinski *et al.* (2017)) is used to establish the tracking baselines on *eTraM*. This technique tracks objects by evaluating the intersection over union values of the bounding boxes detected across sequential frames by an object detector model. The detection baseline results from the Recurrent Event-based Detector (RED) model have been used for the evaluation. Consequently, this method results in a Multi-Object Tracking Precision (MOTP) value of 0.18 and a Multi-Object Tracking Accuracy (MOTA) value of 0.28 on *eTraM*'s test set. It is worth reiterating that the precise evaluation of tracking performance on event data is made possible solely through the inclusion of track-

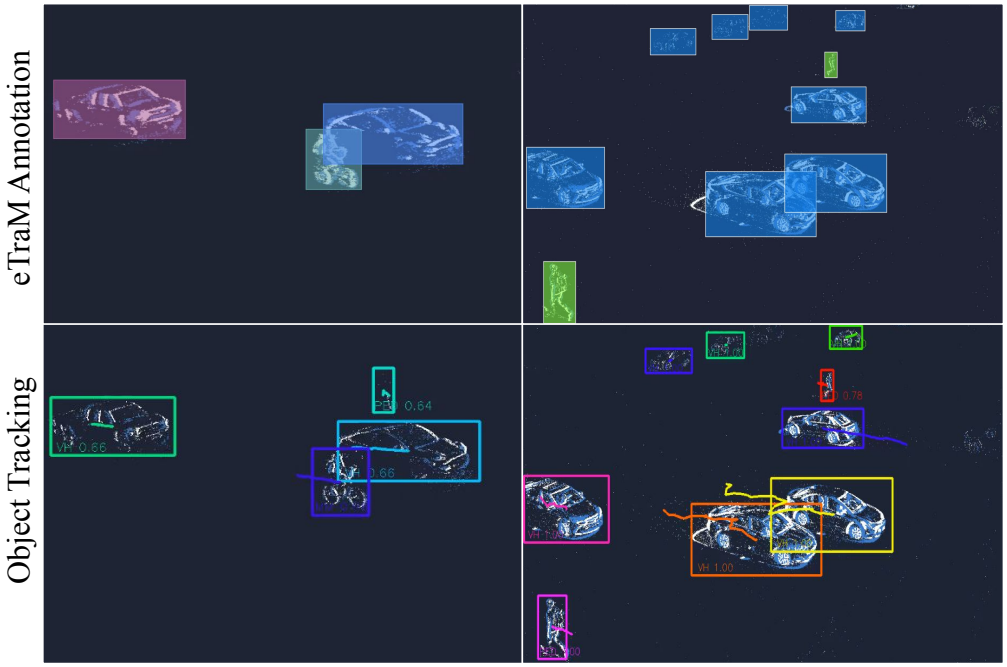

Figure 4.3: Qualitative Results: Showcasing Ground Truth (Top Row) Annotations In *ETraM* And The Corresponding Tracking Of Each Detected Object (Bottom Row) Where Each Trailing Line Denotes The Path Followed By The Detected Object In Previous Timesteps.

ing IDs within *eTraM*. An example of ground truth objects and their corresponding tracking is illustrated in Figure 4.3.

4.3 Impact of Object Size on Detection Performance

To delve further into the impact of object size on model performance across various classes, an additional experiment was conducted, building upon the categorization outlined in the preceding chapter. Objects were classified into three distinct groups based on the area they occupy: small, medium, and large. This finer granularity in classification aimed to provide deeper insights into how the size of objects influences the efficacy of the model across different classes.

Object Size	RVT				RED			
	PED	**VH**	**MM**	**All**	**PED**	**VH**	**MM**	**All**
Small	0.308	0.705	**0.276**	0.516	0.324	0.556	**0.274**	0.385
Medium	**0.859**	**0.722**	0.100	**0.722**	**0.661**	**0.763**	0.159	0.561
Large	-	0.637	-	0.637	-	0.701	-	**0.701**

Table 4.2: Evaluation Of Object Size Impact On The Performance Of RVT And RED.

The established benchmarks resulting from this grouping are presented in Table 4.2. Upon analysis, it becomes evident that both models exhibit similar trends in performance. Specifically, instances categorized as medium-sized within the pedestrian and vehicle classes consistently demonstrate superior performance compared to their smaller and larger counterparts. Although vehicles tend to demonstrate similar performance across all three size classifications, the pedestrian class observes a significant drop in small-sized instances. On the other hand, micro-mobility performs better for small than medium-sized instances. However, the results of micro-mobility in its best-performing size classification are still worse than the worst performance of pedestrian and vehicle categories.

Chapter 5

GENERALIZATION CAPABILITIES OF EVENT DATA

A fundamental requirement for the real-world deployment of event-based detectors is their ability to demonstrate transferability to unseen scenes. This essential characteristic ensures that the detectors can effectively adapt to new environments and scenarios encountered in practical applications. Moreover, given that event cameras are inherently invariant to absolute illuminance levels, it is theoretically expected for event-based detectors to exhibit transferability to nighttime data as well.

In the upcoming sections, the designed experiments quantitatively evaluate the transferability potential of event-based detectors. Controlling the training and test sets in a systematic manner provides comprehensive insights into the extent to which these detectors can generalize across diverse scenes and lighting conditions. Through this set of evaluations, the aim is to ascertain the robustness and adaptability of event-based detectors in real-world scenarios and varied lighting conditions.

Overall, this chapter serves as a detailed exploration into the generalization capabilities of event data, shedding light on their potential for real-world applications and providing valuable insights for future research and development.

5.1 Generalization on Night time

Qualitative assessments of event-based detectors' generalization capabilities to nighttime scenarios were explored in Perot *et al.* (2020), where they discuss how event-based detectors perform better than RGB frame-based detectors during low light conditions. However, since they do not provide annotated nighttime data due to their automated labeling protocol, they only measure the hypothesis qualitatively.

Since *eTraM* consists of night-time annotated data as well, the following experiment aims to quantitatively assess how well a detector trained on daytime data can perform in nighttime conditions and thereby establish the need for night-time data during the training process. A controlled experiment is conducted where an event-based detector is trained on a dataset containing 2 hours of daytime data only. To compare the results of this experiment, the model is fine-tuned with 45 minutes of extra nighttime data. The two models are evaluated on the same set of previously unseen nighttime data sequences.

Train Set	**RVT**		**RED**	
	VH	**PED**	**VH**	**PED**
Day	0.566	0.166	0.374	0.354
Day+Night	0.761	0.254	0.673	0.422

Table 5.1: Evaluation Of Generalization Capabilities Of RED And RVT On Night Time Data For PED - Pedestrian And VH - Vehicle Class For Models Trained On Only Daytime And A Combination Of Daytime And Nighttime Data.

The summarized results of these experiments can be found in Table 5.1. The trend observed is that across every object class and for both detectors, the models trained on data supplemented with nighttime sequences consistently outperform models trained solely on daytime sequences. Despite the expectation that event cameras would exhibit proficient performance in nighttime scenarios, these observations reveal that relying solely on daytime data for training event-based detectors may not be sufficient. These models fail to attain comparable performance levels to those achieved by models trained on nighttime data with the model. This suggests that the unique

challenges posed by nighttime conditions necessitate explicit training with relevant data to ensure optimal detector performance. This discrepancy in performance may be attributed to the unique challenges posed by nighttime conditions, including environmental interferences and distinct variations of noise inherent in nighttime data due to the heightened sensitivity of event-based cameras.

While event-detectors aren't able to generalize on nighttime out of the box, they have at par results on daytime and nighttime data when trained on the combined dataset (Table 4.1). This highlights the need for labeled nighttime data to allow event-based detectors to augment current camera-based systems and perform well during low-light conditions as well.

5.2 Generalization on Unseen Scenes

To substantiate the ability of event-based detectors to generalize to previously unencountered traffic scenes, the model is trained on a subset of the dataset. The ability to generalize is evaluated by testing each architecture on two independent test sets, one "held in" that contains sequences from intersections that the model has seen during the training phase and the other on the "held out" test set with data from an intersection that is skipped during training.

As depicted in Table 5.2, a comparable performance of the models is observed across both the major categories considered. On evaluating the model on the "held out" test set, representing an entirely new and unseen traffic scene, the model's generalization capability is evident. These values are at par with the performance on the "held in" test set. This similarity in results highlights the model's ability to seamlessly extend its learned features and representations to previously unseen traffic scenarios, validating its transferability capability across changing environments.

Test Set	RVT		RED	
	VH	**PED**	**VH**	**PED**
Held In	0.449	0.316	0.556	0.521
Held Out	0.628	0.529	0.572	0.509

Table 5.2: Evaluation Of Generalization Capabilities Of RED And RVT On Unseen Traffic Scenarios For PED - Pedestrian And VH - Vehicle Tested On Held In And Held Out Test Set.

This generalization to unseen intersections is a pivotal characteristic for real-world deployment. The ability to adapt seamlessly to unseen intersections underscores its potential applicability in ITS and reinforces its suitability for a broad range of applications.

Chapter 6

DISCUSSION

This chapter aims to explore and start a discussion on the potential applications and advantages of utilizing fixed event-based cameras for traffic monitoring, particularly in static roadside environments. While much of the current research in the field predominantly focuses on ego-motion cameras, this discussion sheds light on the unique benefits that fixed event cameras offer and the various scenarios where they could be leveraged effectively.

6.1 Event Cameras in Traffic Monitoring

In this section, the advantages of utilizing event cameras to augment traditional frame camera systems for various traffic monitoring tasks in static roadside environments are explored.

Traditional cameras take continuous snapshots at a fixed frequency, potentially capturing individuals' identities and sensitive information. In comparison, event cameras register events significantly, reducing the probability of gathering sensitive information of any individual Perot *et al.* (2020); Bolten *et al.* (2021). The advantages of event cameras extend to their performance in low light conditions and their robustness towards mitigating motion blur. As demonstrated through the baseline evaluations in Table 4.1, event cameras display equally superior performance in nighttime conditions while maintaining a high temporal resolution that aids it in substantially reducing motion blur Rebecq *et al.* (2021). An essential requirement of long-time traffic monitoring is the need for resource-efficient sensors. Since event cameras are designed to operate on a sparse data stream generated by significant visual changes, they ex-

hibit lower memory requirements and power consumption compared to traditional frame cameras Amir *et al.* (2017); Serrano-Gotarredona *et al.* (2009). This makes event cameras sustainable and cost-effective for continuous monitoring over extended periods.

Figure 6.1: Demonstrating Effectiveness Of Event Camera For Traffic Scenarios: Yellow Circle (Top Row) Tracks A Car That Halts At A Stop Sign With Lack Of Motion Captured In The Third Frame, Red Circle (Bottom Row) Tracks A Car That Violates The Stop Sign Where Motion Is Continuously Captured In Every Frame. Additionally, The Green Arrow (Top Row) Shows A Car Traveling At A High Speed, Resulting In A High Event Density.

The effectiveness of event cameras in traffic monitoring scenarios is further highlighted through two practical applications. Firstly, in detecting stop sign violations, event cameras excel in capturing instantaneous changes in the visual scene, enabling precise detection of vehicles coming to a halt. Secondly, event cameras are adept at quickly detecting sudden acceleration or erratic behavior of fast-moving traffic participants. This is a critical capability for identifying potentially hazardous situations on the road and allowing for prompt intervention or alert generation. Figure 6.1 provides a clear indicator - the number of events associated with the object - making it easy to discern whether a moving vehicle is slowing down or has stopped. Given the effectiveness demonstrated by event cameras in specific traffic monitoring scenarios,

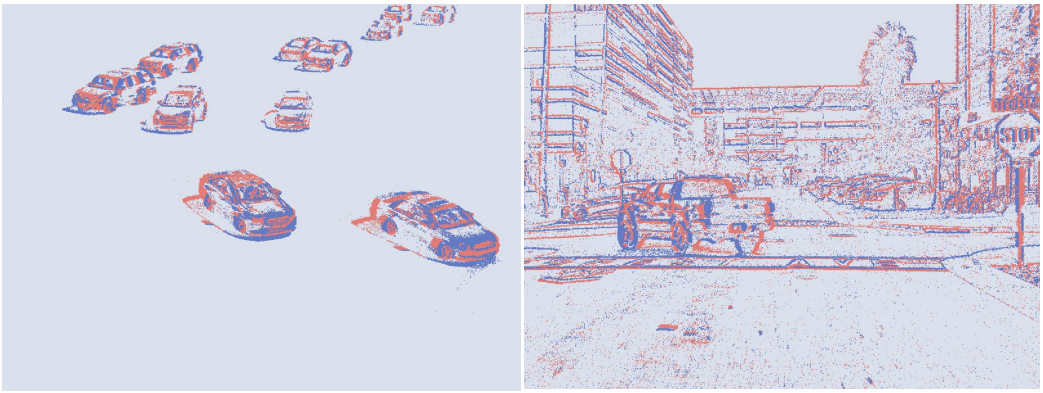

Figure 6.2: Qualitative Comparison: Events Captured From A Static Event Camera (Left) Show Enhanced Visibility Of Moving Vehicles Compared To An Ego-motion Event Camera (Right).

such as detecting stop sign violations and identifying sudden acceleration or erratic behavior, it is evident that further explorations in multiview event-based traffic monitoring (Aliminati *et al.* (2024))are warranted. Such systems have the potential to improve the robustness and reliability of traffic monitoring systems by reducing blind spots, mitigating occlusions, and improving tracking performance.

6.2 Static and Ego Event-based Datasets

Here, the difference between static event-based datasets and existing ego-motion event-based datasets is discussed, emphasizing the significance of the former for traffic participant detection. A notable disparity between event data and traditional frame-based data lies in the sparsity of relevant information. This distinct feature allows event cameras to operate with sub-millisecond latency without compromising resolution due to bandwidth limitations, as often observed in frame-based cameras. However, ego-motion datasets tend to sacrifice sparsity in event data because the entire background moves relative to the camera, resulting in events being generated across

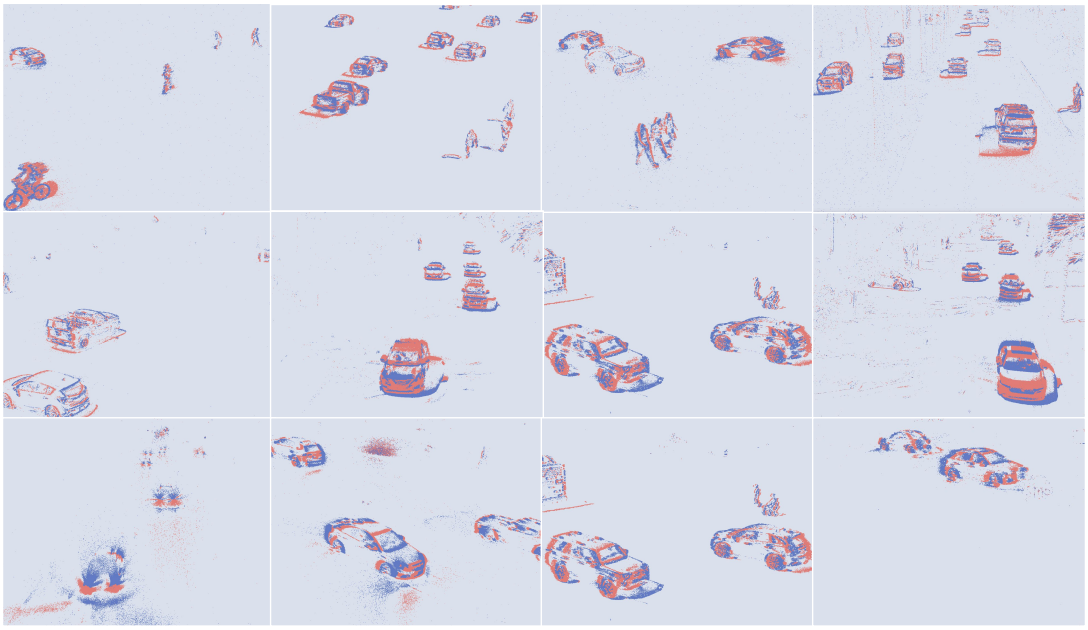

Figure 6.3: Traffic Site Diversity In *ETraM*: Various Instances Encapsulating The Interactions Amongst Multiple Traffic Participants Captured From A Static Roadside Perspective Are Shown With Daytime (First Row), Twilight (Second Row), And Nighttime (Last Row) Showing Increasing Sensor Noise (Top To Bottom) Due To Light Sources Such As Headlights And Streetlights.

most of the frame. In contrast, static perspective cameras capitalize on the sparse nature of event cameras, where events are more likely to correspond to relevant traffic entities rather than extraneous background changes. This characteristic significantly enhances the visibility of traffic participants too, as shown in Figure 6.2. Another distinction is the field of view provided by static roadside datasets. Positioned at an elevated height, these cameras capture a broader scene that includes far-away objects, while ego-motion datasets are constrained to the vehicle's immediate surroundings. The elevated perspective of static datasets proves advantageous, enabling the detection of traffic participants at a distance and providing a comprehensive view of the traffic environment, as seen in Figure 6.3.

Chapter 7

CONCLUSION

This chapter summarises the thesis, the limitations of the work, and a few future directions in event-based research.

7.1 Summary

The thesis proposes *eTraM*, a large real-world, manually annotated event-based dataset for event-based traffic monitoring from a static perspective. The meticulously curated dataset, captured using the cutting-edge IMX636HD (Prophesee EVKHD4) high-resolution event camera, provides new opportunities in the world of traffic detection and tracking from a static roadside perspective. With over 10 hours of annotated event data spanning various environmental conditions, *eTraM* offers insights into the complex dynamics of traffic scenarios. The comprehensive annotations, which include over 2 million bounding boxes of various traffic participants, from vehicles to pedestrians and micro-mobility, allow for a holistic understanding of the challenges posed by diverse scenarios and participants. Through various experiments, the ability of event-based models to generalize effectively to unseen scenarios is demonstrated, thus emphasizing its potential in real-world traffic monitoring. Nighttime generalization experiment results highlight the value of labeled nighttime data, enabling the same event-based detectors to achieve performance that is competitive with daytime performance. As the field of event-based sensing continues to advance, and as the demand for enhanced traffic safety within intelligent transportation systems grows, *eTraM* holds the potential to serve as an invaluable resource that can drive the development of road safety and traffic management.

7.2 Limitations and Future Work

Despite the significant contributions of *eTraM* and the research conducted, there are several limitations and future work to improve the current state-of-the-art systems:

1. Conversion to Image/Tensor Representations: The current approaches explored involve converting event data into image or tensor representations, thereby neglecting the inherent sparsity and temporal resolution in *eTraM*. Future research could focus on developing techniques leveraging spiking neural networks and graph-based representations that leverage the unique characteristics of event-based data without resorting to conventional image-based representations.

2. Performance on Small Objects: Event-based detectors exhibit notable performance degradation when dealing with small-sized objects, particularly micromobility. This limitation may stem from the lack of contour and color information in raw event data. Addressing this challenge could involve leveraging other sensors to complement event data.

3. Class Imbalance in Micromobility Instances: The imbalance in the number of instances between micromobility and vehicles is observed due to the inherent dynamics of the traffic environment. Various data augmentation techniques and making use of large amounts of synthetic data could help alleviate class imbalance issues and improve model generalization.

4. Expansion of Dataset: Expanding *eTraM* to encompass recordings from different geographical locations and incorporating additional annotation tasks beyond traffic monitoring, such as anomaly detection, could broaden the dataset's scope and applicability.

REFERENCES

"Event Camera Evaluation Kit 4 HD IMX636 Prophesee-Sony", URL https://www.prophesee.ai/event-camera-evk4/ (2023).

Aliminati, M. R., B. Chakravarthi, A. A. Verma, A. Vaghela, H. Wei, X. Zhou and Y. Yang, "Sevd: Synthetic event-based vision dataset for ego and fixed traffic perception", arXiv preprint arXiv:2404.10540 (2024).

Amir, A., B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner and D. Modha, "A low power, fully event-based gesture recognition system", in "2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", pp. 7388–7397 (2017).

Baldwin, R. W., R. Liu, M. Almatrafi, V. Asari and K. Hirakawa, "Time-ordered recent event (tore) volumes for event cameras", IEEE Transactions on Pattern Analysis and Machine Intelligence **45**, 2, 2519–2532 (2022).

Binas, J., D. Neil, S.-C. Liu and T. Delbruck, "Ddd17: End-to-end davis driving dataset", (2017).

Bochinski, E., V. Eiselein and T. Sikora, "High-speed tracking-by-detection without using image information", in "International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017", (Lecce, Italy, 2017), URL http://elvera.nue.tu-berlin.de/files/1517Bochinski2017.pdf.

Bolten, T., R. Pohle-Fröhlich and K. D. Tönnies, "Dvs-outlab: A neuromorphic event-based long time monitoring dataset for real-world outdoor scenarios", in "2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)", pp. 1348–1357 (2021).

Boretti, C., P. Bich, F. Pareschi, L. Prono, R. Rovatti and G. Setti, "Pedro: An event-based dataset for person detection in robotics", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 4064–4069 (2023).

Brosch, T., S. Tschechne and H. Neumann, "On event-based optical flow detection", Frontiers in neuroscience **9**, 137 (2015).

Caesar, H., V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", (2020).

Chen, G., H. Cao, J. Conradt, H. Tang, F. Rohrbein and A. Knoll, "Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception", IEEE Signal Processing Magazine **37**, 4, 34–49 (2020).

Dalgaty, T., T. Mesquida, D. Joubert, A. Sironi, P. Vivet and C. Posch, "Hugnet: Hemi-spherical update graph neural network applied to low-latency event-based optical flow", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 3952–3961 (2023).

de Tournemire, P., D. Nitti, E. Perot, D. Migliore and A. Sironi, "A large scale event-based detection dataset for automotive", (2020).

Dosovitskiy, A., G. Ros, F. Codevilla, A. Lopez and V. Koltun, "CARLA: An open urban driving simulator", in "Proceedings of the 1st Annual Conference on Robot Learning", pp. 1–16 (2017).

Fei-Fei, L., R. Fergus and P. Perona, "One-shot learning of object categories", IEEE Transactions on Pattern Analysis and Machine Intelligence **28**, 4, 594–611 (2006).

Gallego, G., T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis and D. Scaramuzza, "Event-based vision: A survey", CoRR **abs/1904.08405**, URL http://arxiv.org/abs/1904.08405 (2019).

Gehrig, D., M. Gehrig, J. Hidalgo-Carrió and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras", in "IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)", (2020).

Gehrig, M., W. Aarents, D. Gehrig and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios", IEEE Robotics and Automation Letters (2021).

Gehrig, M. and D. Scaramuzza, "Recurrent vision transformers for object detection with event cameras", (2023).

Geiger, A., P. Lenz and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite", in "2012 IEEE Conference on Computer Vision and Pattern Recognition", pp. 3354–3361 (2012).

Ghari, B., A. Tourani, A. Shahbahrami and G. Gaydadjiev, "Pedestrian detection in low-light conditions: A comprehensive survey", ArXiv **abs/2401.07801**, URL https://api.semanticscholar.org/CorpusID:266998730 (2024).

Harley, A. W., Z. Fang, J. Li, R. Ambrus and K. Fragkiadaki, "Simple-BEV: What really matters for multi-sensor bev perception?", in "arXiv:2206.07959", (2022).

Hu, Y., J. Binas, D. Neil, S.-C. Liu and T. Delbruck, "Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction", (2020).

Hu, Y., S.-C. Liu and T. Delbruck, "v2e: From video frames to realistic dvs events", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops", pp. 1312–1321 (2021).

Jocher, G., A. Chaurasia and J. Qiu, "YOLO by Ultralytics", URL https://github.com/ultralytics/ultralytics (2023).

Lagorce, X., G. Orchard, F. Galluppi, B. E. Shi and R. B. Benosman, "Hots: A hierarchy of event-based time-surfaces for pattern recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 7, 1346–1359 (2017).

Lecun, Y., L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE **86**, 11, 2278–2324 (1998).

Liu, M. and T. Delbrück, "Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors", in "British Machine Vision Conference", (2018), URL https://api.semanticscholar.org/CorpusID:52283776.

Liu, M. and T. Delbruck, "Edflow: Event driven optical flow camera with keypoint detection and adaptive block matching", IEEE Transactions on Circuits and Systems for Video Technology **32**, 9, 5776–5789 (2022).

Maqueda, A. M. I., A. Loquercio, G. Gallego, N. García and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 5419–5427, URL https://api.semanticscholar.org/CorpusID:4610262 (2018).

Mesquida, T., M. Dampfhoffer, T. Dalgaty, P. Vivet, A. Sironi and C. Posch, "G2N2: Lightweight event stream classification with GRU graph neural networks", in "BMVC 2023 - The 34th British Machine Vision Conference", https://proceedings.bmvc2023.org/, p. 660 (Aberdeen, United Kingdom, 2023), URL https://cea.hal.science/cea-04321175.

Messikommer, N., D. Gehrig, A. Loquercio and D. Scaramuzza, "Event-based asynchronous sparse convolutional networks", (2020), URL http://rpg.ifi.uzh.ch/docs/ECCV20_Messikommer.pdf.

Miao, S., G. Chen, X. Ning, Y. Zi, K. Ren, Z. Bing and A. Knoll, "Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection", Frontiers in Neurorobotics **13**, URL https://www.frontiersin.org/articles/10.3389/fnbot.2019.00038 (2019).

Mueggler, E., B. Huber and D. Scaramuzza, "Event-based, 6-dof pose tracking for high-speed maneuvers", in "2014 IEEE/RSJ International Conference on Intelligent Robots and Systems", pp. 2761–2768 (2014).

Orchard, G., A. Jayawant, G. Cohen and N. Thakor, "Converting static image datasets to spiking neuromorphic datasets using saccades", (2015).

Perot, E., P. de Tournemire, D. Nitti, J. Masci and A. Sironi, "Learning to detect objects with a 1 megapixel event camera", in "Advances in Neural Information Processing Systems", edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, vol. 33, pp. 16639–16652 (Curran Associates, Inc., 2020), URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c213877427b46fa96cff6c39e837ccee-Paper.pdf.

Prophesee, "Prophesee - Metavision for Machines", URL https://www.prophesee.ai/ (2023).

Prophesee S.A, "Prophesee Documentation - Event-Based Concepts", URL https://docs.prophesee.ai/stable/concepts.html, © Copyright Prophesee S.A - All Rights Reserved. (2024a).

Prophesee S.A, "Prophesee Documentation - File Formats", URL https://docs.prophesee.ai/stable/data/file_formats/index.html, © Copyright Prophesee S.A - All Rights Reserved. (2024b).

Rebecq, H., D. Gehrig and D. Scaramuzza, "Esim: an open event camera simulator", in "Proceedings of The 2nd Conference on Robot Learning", edited by A. Billard, A. Dragan, J. Peters and J. Morimoto, vol. 87 of *Proceedings of Machine Learning Research*, pp. 969–982 (PMLR, 2018), URL https://proceedings.mlr.press/v87/rebecq18a.html.

Rebecq, H., R. Ranftl, V. Koltun and D. Scaramuzza, "High speed and high dynamic range video with an event camera", IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 6, 1964–1980 (2021).

Rodríguez-Gomez, J., A. G. Eguíluz, J. Martínez-de Dios and A. Ollero, "Asynchronous event-based clustering and tracking for intrusion monitoring in uas", in "2020 IEEE International Conference on Robotics and Automation (ICRA)", pp. 8518–8524 (2020).

Schaefer, S., D. Gehrig and D. Scaramuzza, "Aegnn: Asynchronous event-based graph neural networks", in "IEEE Conference on Computer Vision and Pattern Recognition", (2022).

Sekachev, B., N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, TOsmanov, D. Kruchinin, A. Zankevich, DmitriySidnev, M. Markelov, Johannes222, M. Chenuet, a andre, telenachos, A. Melnikov, J. Kim, L. Ilouz, N. Glazov, Priya4607, R. Tehrani, S. Jeong, V. Skubriev, S. Yonekura, vugia truong, zliang7, lizhming and T. Truong, "opencv/cvat: v1.1.0", URL https://doi.org/10.5281/zenodo.4009388 (2020).

Serrano-Gotarredona, R., M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gomez-Rodriguez, L. Camunas-Mesa, R. Berner, M. Rivas-Perez, T. Delbruck, S.-C. Liu, R. Douglas, P. Hafliger, G. Jimenez-Moreno, A. Civit Ballcels, T. Serrano-Gotarredona, A. J. Acosta-Jimenez and B. Linares-Barranco, "Caviar: A 45k neuron, 5m synapse, 12g connects/s aer hardware sensory–processing– learning–actuating system for high-speed visual object recognition and tracking", IEEE Transactions on Neural Networks **20**, 9, 1417–1438 (2009).

Sironi, A., M. Brambilla, N. Bourdis, X. Lagorce and R. B. Benosman, "Hats: Histograms of averaged time surfaces for robust event-based object classification", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 1731–1740, URL https://api.semanticscholar.org/CorpusID:3993392 (2018).

Sun, P., H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen

and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset", CoRR **abs/1912.04838**, URL http://arxiv.org/abs/1912.04838 (2019).

Sussman, J. S., *Perspectives on intelligent transportation systems (ITS)* (Springer Science & Business Media, 2008).

Tomy, A., A. Paigwar, K. S. Mann, A. Renzaglia and C. Laugier, "Fusing event-based and rgb camera for robust object detection in adverse conditions", in "2022 International Conference on Robotics and Automation (ICRA)", pp. 933–939 (2022).

Verma, A. A., B. Chakravarthi, A. Vaghela, H. Wei and Y. Yang, "etram: Event-based traffic monitoring dataset", arXiv preprint arXiv:2403.19976 (2024).

Wang, L., I. M. Mostafavi, Y.-S. Ho and K.-J. Yoon, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks", in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", pp. 10073–10082 (2019).

Yang, Z. and L. S. Pun-Cheng, "Vehicle detection in intelligent transportation systems and its applications under varying environments: A review", Image and Vision Computing **69**, 143–154 (2018).

Yao, M., H. Gao, G. Zhao, D. Wang, Y. Lin, Z. Yang and G. Li, "Temporal-wise attention spiking neural networks for event streams classification", in "2021 IEEE/CVF International Conference on Computer Vision (ICCV)", pp. 10201–10210 (IEEE Computer Society, Los Alamitos, CA, USA, 2021), URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01006.

Zhang, J., B. Dong, H. Zhang, J. Ding, F. Heide, B. Yin and X. Yang, "Spiking transformers for event-based single object tracking", in "2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", pp. 8791–8800 (2022).

Zhang, Z., L. Zhang, D. Meng, L. Huang, W. Xiao and W. Tian, "Vehicle kinematics-based image augmentation against motion blur for object detectors", Tech. rep., SAE Technical Paper (2023).

Zheng, X., Y. Liu, Y. Lu, T. Hua, T. Pan, W. Zhang, D. Tao and L. Wang, "Deep learning for event-based vision: A comprehensive survey and benchmarks", (2023).

Zhu, A. Z., D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception", IEEE Robotics and Automation Letters **3**, 3, 2032–2039 (2018a).

Zhu, A. Z., L. Yuan, K. Chaney and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 989–997, URL https://api.semanticscholar.org/CorpusID:56475917 (2018b).