Supporting the Intersection Safety Project

by

Jayaram Atluri

Submitted to:
Dr. Jeffrey Wishart

ARIZONA STATE UNIVERSITY

December 2023

ABSTRACT

This project aimed to develop an intersection safety system using event cameras and AI for multi-sensor fusion. The primary objectives included setting up and calibrating event cameras, collecting and synchronizing sensor data streams, processing, and labelling the data, conducting research to enable sensor fusion techniques, and providing engineering support for overall model development. While working on the project, several challenges were faced such as handling asynchronous event data, achieving accurate alignment between event and RGB frames, and transforming irregular event streams into grid-like tensors.

To address these challenges, several innovative techniques such as projection, accumulation, and encoder-decoder networks were suggested. The event data was encoded into spatially organized representations using these techniques, enabling fusion with RGB frames using standard computer vision approaches. A two-stage fused model was suggested that first generates region proposals combining motion patterns from events and spatial context from RGB, before classifying objects. Optical flow and LSTM networks were researched to enable precise tracking, motion forecasting, and anomaly detection for vehicles and vulnerable road users at intersections.

Throughout the project, significant learnings were gained regarding optimal sensor selection, multi-modal sensor fusion architectures, and end-to-end development of perception for autonomous vehicles. However, there remains a need for further progress in areas such as domain-specific modelling, robustness testing, and real-world validation of the safety system. The project findings underscore the complexity and depth of considerations necessary for the successful integration of event cameras and AI in intersection safety systems, offering a valuable foundation for future advancements in this critical domain.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS/NOMENCLATURE

Symbols/Nomenclature

1. AI - Artificial Intelligence
2. ASU - Arizona State University
3. CAV - Connected and Automated Vehicles
4. CVAT - Computer Vision Annotation Tool
5. DOT - Department of Transportation
6. EVK4 - Event Camera Evaluation Kit 4
7. FHWA-JPO - Federal Highway Administration Joint Program Office
8. ISS - Intersection Safety System
9. LSTM - Long Short-Term Memory
10. ML - Machine Learning
11. PASCAL VOC - Pattern Analysis, Statistical Modeling, and Computational Learning of Visual Object Classes
12. RGB - Red, Green, Blue
13. RFI - Request for Information
14. TF Record - TensorFlow Record
15. USDOT - U.S. Department of Transportation

1 INTRODUCTION

Improving safety at intersections is a major priority for enabling large-scale deployment of connected and automated vehicles (CAVs). Intersections witness a high number of crashes due to conflicting points between vehicles, pedestrians, and cyclists. Developing intelligent safety technologies to understand, predict and communicate all elements within the intersection ecology is crucial for avoiding incidents and protecting vulnerable road users.

This report provides a detailed overview of my semester-long project undertaken under the sponsorship of Dr. Yezhou Yang's Active Perception Group and lab at Computing and Augmented Intelligence, Arizona State University. The lab is currently working on an 'Intersection Safety Challenge project sponsored by USDOT, aiming to create a next-generation safety system using sensing and perception technologies. As part of this initiative, the project documented in this report focused on investigating event-based vision sensor cameras and their fusion with conventional RGB cameras for reliable environment perception especially at intersections.

1.1 Overview of the Intersection Safety Challenge

The Intersection Safety Challenge is an effort launched by the U.S. Department of Transportation (DOT) to spur innovation around intersection safety, especially for vulnerable road users. Intersections are high-risk areas, accounting for 27% of all traffic fatalities. Pedestrian and bicyclist fatalities are also rising based on recent data.

The DOT aims to facilitate the development of an Intersection Safety System (ISS) that leverages emerging technologies to identify and mitigate unsafe conditions at intersections in real time. The vision is to transform intersection safety through systems that can detect all intersection users, predict movements, issue multimodal warnings, and modify traffic signals or other controls to prevent collisions (Figure 1).

Key capabilities of the ISS include simultaneous road user detection/tracking, trajectory prediction, wireless connectivity, integration with legacy systems, interoperability between neighbouring intersections, issuance of warnings to vehicles and unconnected vulnerable users, and upgrades over time.

The DOT is utilizing a prize competition model (Figure 2) to incentivize collaborations between industry, academia, and public agencies to conceptualize, prototype, and field-test ISS solutions. Winning concepts must demonstrate life-saving potential combined with cost-effectiveness and a viable path to mainstream adoption within 10 years.

Our team recognizes the Intersection Safety Challenge as an unparalleled opportunity to contribute our expertise in connected and automated vehicles, embedded systems, and human-computer interaction to this game-changing innovation program aligned with our own mission of safer and more equitable transportation through technology and policy co-optimization.

Figure 1: Concept Illustration: Intersection Safety System
Safety systems informed by data fused from multiple sensors may anticipate unsafe conditions, e.g., a vehicle turning right in potential conflict with a pedestrian pushing a stroller.
Image Source: U.S. DOT.

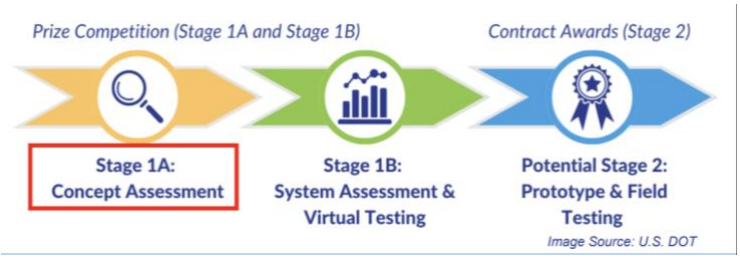


Figure 2: Program Structure/Prize competition model

1.2 The Intersection Safety System (ISS) concept

The Intersection Safety System (ISS) is an innovative concept to improve safety at intersections using emerging technologies. The core idea is to deploy low-cost sensors like cameras, radar, LiDAR, and infrared at intersections to better detect different types of road users including vehicles, pedestrians, bicyclists, and other micro-mobility devices. Advanced software and artificial intelligence would then fuse the sensor data and analyze it in real-time to anticipate potential crashes and conflicts. The system could then issue warnings to vehicles and road users and adapt traffic signals or other controls to prevent collisions.

Key technical capabilities of the ISS include simultaneous detection, localization and classification of all road users, prediction of vehicle and vulnerable road user movements, data handling and storage, wireless communications and positioning, interaction with existing traffic control systems, issuance of warnings, and interoperability between intersections (Figure 3).

The system must maintain reliable performance, provide a cost-effective solution, have a path to commercialization and deployment within 10 years, and allow for upgradeability and modularity.
The vision for the ISS is the transformation of intersection safety through innovative systems that identify, predict and mitigate unsafe conditions in real-time involving vehicles and vulnerable road users. The ISS concept aligns with my project's focus on applying technology innovations to improve transportation safety and accessibility outcomes. Research and development of ISS prototypes could ultimately lead to lifesaving solutions that support the National Roadway Safety Strategy goal of zero deaths and serious injuries.

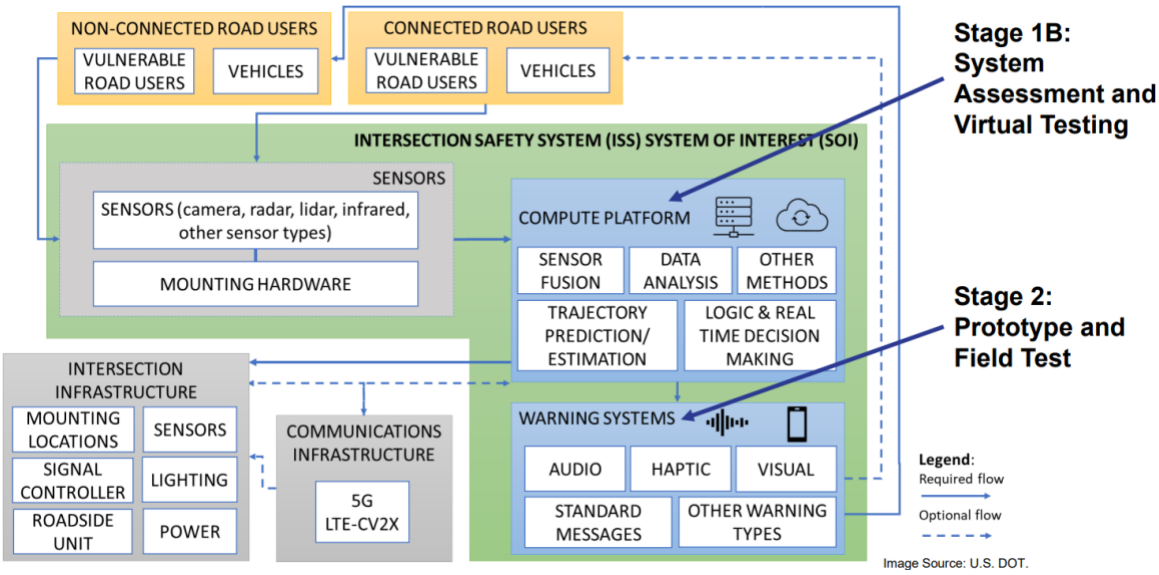


Figure 3: ISS Concept

1.3 Project Problem Definition

The main issue addressed in this project is how to capture all the different moving components within busy urban intersections with the help of a perfect sensor suite. The data collected will be intelligently processed to identify traffic participants and model their movements. To achieve this goal, we need to tackle several sub-problems such as:

1. Dealing with variable lighting, occlusion, and varying densities of vehicles and pedestrians
2. Balancing sensor resolution, range, field-of-view, and data bandwidth
3. Synchronizing asynchronous event and RGB streams with minimum latency
4. Transforming irregular event data into spatially aligned representations.
5. Designing multi-modal neural networks for detection and tracking
6. Ensuring model generalizability across locations, times, and events

1.4 Project Objectives

The main objectives of this project are:

1. Setting up and calibrating event-based vision cameras for data collection
2. Recording synchronized event camera and RGB video feeds at intersections
3. Investigating algorithms for sensor alignment and data transformation
4. Researching techniques to combine event and RGB streams.
5. Providing engineering support for overall model development
6. Documenting key aspects to aid future teams working in this domain.

Please note that the project does not involve physical infrastructure development, hardware, network deployment, integration with existing systems, or environmental impact. The focus is solely on the development and deployment of the AI-based intersection safety system and its associated components. The subsequent sections discuss the approaches adopted and outcomes achieved while addressing the above goals over 11 weeks.

2 PROJECT REVIEW

As part of our project study, we thoroughly researched various aspects of traffic detection and tracking at intersections. This included a detailed analysis of sensor hardware, calibration methods, and fusion techniques. We also investigated domain-specific models that are currently being used for traffic participant detection and tracking. Based on our findings, we made suggestions for improvement and implemented them where necessary. We aimed to achieve a more accurate and efficient system for traffic detection and tracking.

2.1 Event-based Vision Sensors

Event cameras, like the Prophesee Metavision sensor (Figure 4), operate differently from traditional cameras that use image sensors. As explained in [1], these cameras detect changes in intensity at the microsecond resolution level, rather than capturing images at a fixed rate (Figure 5). This means that they can capture very fast motions and high dynamic ranges, which makes them well-suited for autonomous driving scenarios.

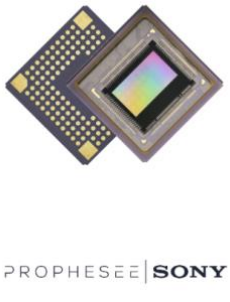


Figure 4: Metavision® sensor IMX636

Figure 5: Representation of synchronous frames of RGB camera versus asynchronous pixels of an event camera

However, processing event data requires new algorithms because it is irregular. Initially, researchers attempted to accumulate events over time and encode them as 2D frames. However, with recent advancements in deep learning, end-to-end models have been developed that use raw events directly, as summarized in [2]. This has prompted further investigation into methods that can transform sporadic events into grid-based representations, which can then be fused with standard RGB inputs.

2.2 Multi-modal Sensor Fusion

An intelligent intersection safety system requires the integration of various sensors such as cameras, lidars, and radars. To achieve this, classical and learned techniques are used to fuse these sensors synergistically.

Several fusion methods have been studied, including early fusion methods that concatenate sensor inputs using neural layers (explored in [3]), late fusion approaches that combine deep feature representations (analyzed in [4]), and novel recurrent fusion networks using LSTM encoder-decoder stacks that model temporal event contexts (examined via [5]). These methods provide insights on how to transform irregular event streams into spatially aligned tensors, leveraging both the motion encoding strengths and RGB context.

In addition, specific techniques such as optical flow, object detection and segmentation, motion forecasting, and anomaly detection were also studied. These techniques are essential to connect research insights with the development requirements of this project.

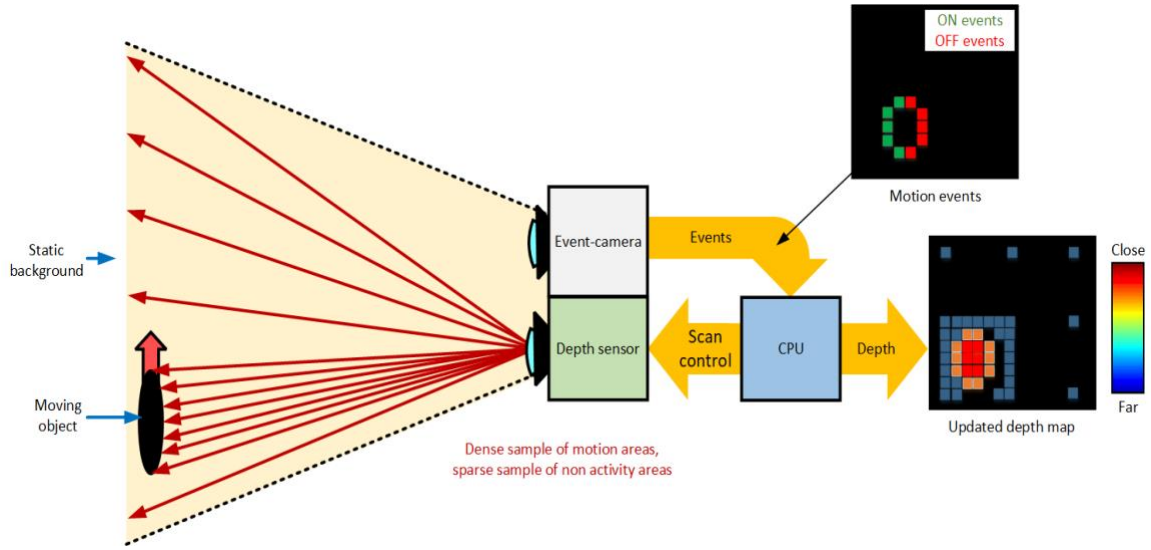


Figure 6: Event-based vision Representation.

2.3 Event Camera, Setup and Data Collection

The initial phase focused on gaining hands-on experience with event cameras and setting up a multi-modal data capture rig.

2.3.1 Event Camera Operation

The Prophesee Gen4 Automotive sensor camera (Figure 7) evaluation kit was utilized in this project. The project involved studying the operation of the camera concerning various metrics, such as pixel contrast thresholds, sensitivity tuning, background noise, and more, as discussed in [6] (Figure 8). This study revealed that achieving a balance between sensitivity to detect motion and excess noise requires a characterization specific to the application.

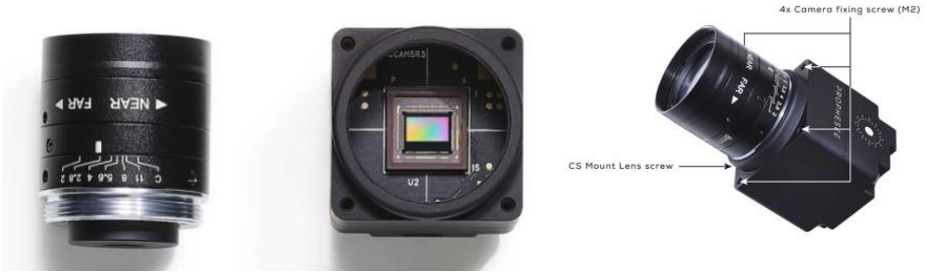


Figure 7: Event Camara (Industry's Smallest*1 4.86µm Pixel Size for Detecting Subject Changes)

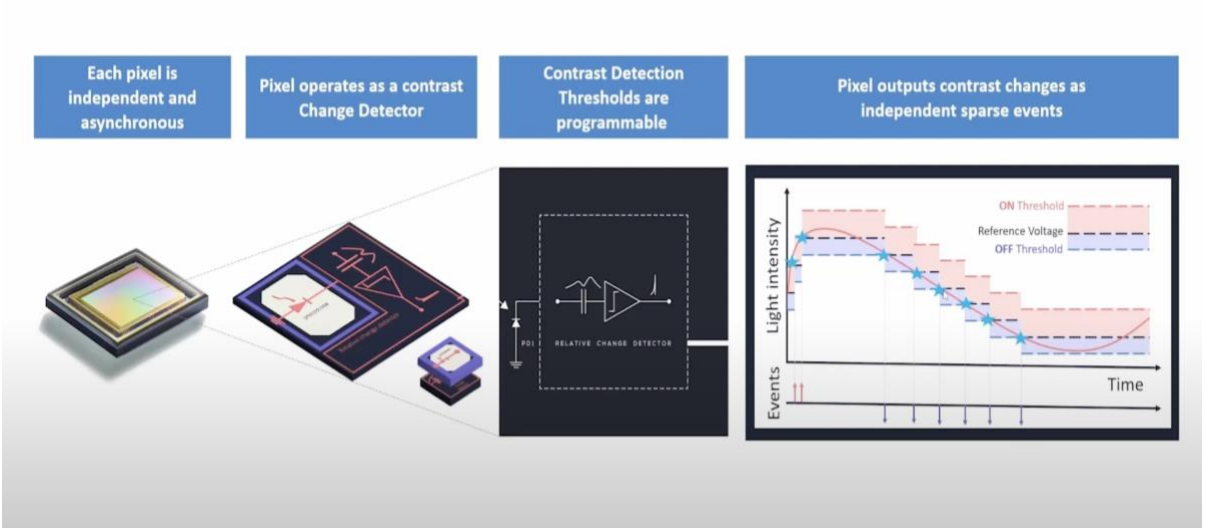


Figure 8: Pixel Architecture

Here are some important key points about event sensors:

1. Event cameras detect changes based on an adjusted threshold (ON - OFF - Ref), generating events based on the set contrast sensitivity threshold.
2. Sensitivity bias can be adjusted based on the intensity change that needs to be detected for the application, generating data only when crossing the contrast sensitivity threshold.
3. It is observed that contrast sensitivity varies with lighting and luminance levels.
4. Higher sensitivity generates more data, so there must be a trade-off between detectable sensitivity and data generation.
5. The background rate refers to the number of events generated over time under constant.

2.3.2 Sensor Calibration

For our project, it was crucial to accurately calibrate and synchronize the event and RGB cameras. We examined several methods described in [7] and ultimately decided to use a black-and-white LCD video (Figure 9) as a cost-effective alternative. This process allowed us to address the challenges of aligning microsecond event exposures with millisecond RGB frames.

To calibrate event cameras, there are a few common methods available. These include using a blinking LED or LCD pattern, an LCD/OLED display, a servo motor rig, or visual-inertial calibration. The blinking LED board method is the most used, but the display-based approach is cheaper. However, it can introduce errors. Servo rig and visual-inertial calibration methods are highly accurate but require complex setups. In this paper [5], it is proposed an image reconstruction method that overcomes previous limitations.

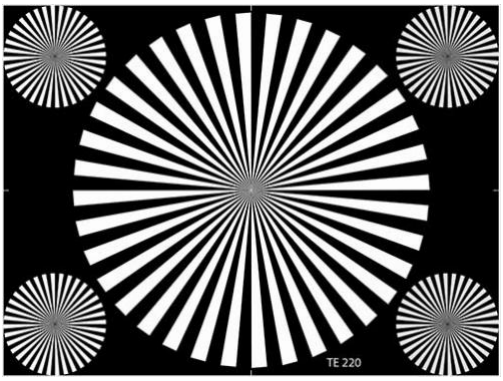


Figure 9: Black-and-white video for setting parameters and calibration

2.3.3 Data Capture Platform

For our experiment, we used a fixed test and camera stand to mount both the RGB and event cameras as close together as possible. We ensured that video capture started simultaneously, but there may have been a mismatch in starting frames, which we corrected by adding a starting frame recognizer video (Figure !0), as shown in the figure below. I proposed this idea of this to identify the starting frame, which is crucial for fusion. Each frame was different; there were no repeated frames.

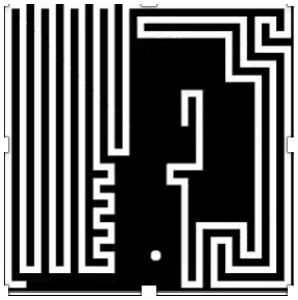


Figure 10: starting frame recognizer Video Snippet

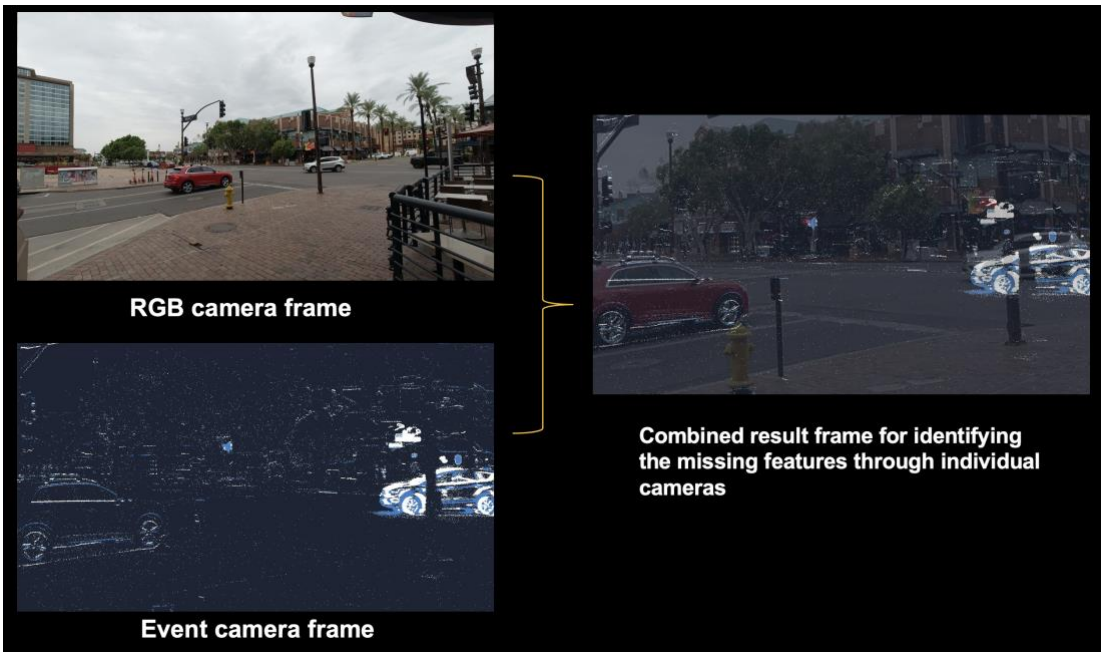


Figure 11: RGB frame versus event frame

Metavision Studio (Figure 12) is an essential tool for working with Prophesee-compatible event-based vision systems. It offers a user-friendly Graphical User Interface that allows us to visualize and record data streamed by these systems. The software comes with RAW tiles in its sample recordings, which, when paired with Evaluation Kits or a compatible camera, makes it easier to adjust display parameters and tune all camera settings for optimal performance.

We used the Metavision Recording Suite with the setup shown in Figure 1 below to record all the vehicle and walking participant footage at intersections. This enabled us to develop across multiple object classes.

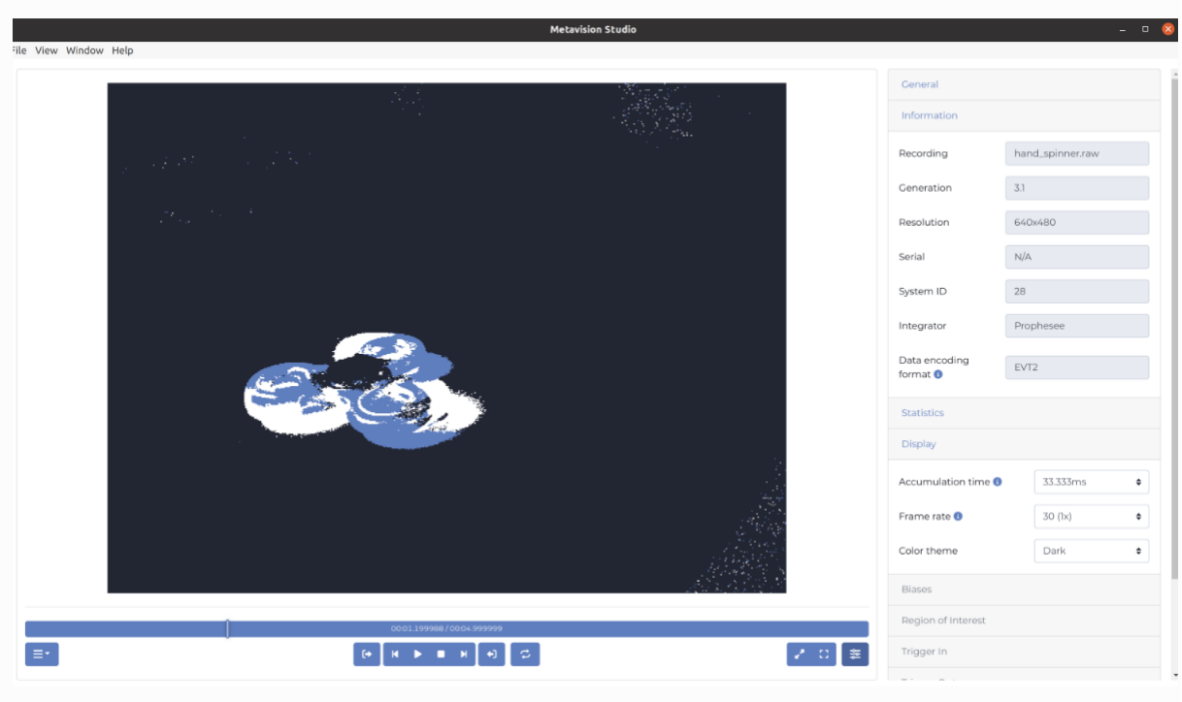


Figure 12: Metavision Studio Tool

2.4 Data Processing and Annotation

We utilized CVAT.ai, an open-source video and image annotation toolbox, to label various traffic participants, such as cars, trucks, pedestrians, and bikes, across thousands of frames. The platform's intuitive bounding box and polygon drawing tools, along with object track auto-propagation, significantly reduced the annotation time compared to frame-by-frame labelling. Additionally, the platform's customizable labels and attributes tailored to our unique traffic participant classes ensured we could capture all necessary metadata.

The platform's detailed analytics on labeller activity and consensus helped us maintain high inter-annotator agreement and catch inconsistencies early on. We were able to export the annotated frames in various formats, such as PASCAL VOC, COCO, TFRecord, and CVAT Dumper, which directly fed into our selected ML frameworks like TensorFlow for training, testing, and validation.

CVAT.ai's cloud-based platform gave our globally distributed team concurrent, remote access to the annotation project. We could also integrate our customized AI-assistance models to semi-automate parts of the repetitive labelling work. Overall, with CVAT's rich feature set, we were able to annotate diverse, large-scale street scene video rapidly and accurately to enable our downstream ML tasks like automated traffic participant tracking and trajectory forecasting. The flexibility of the platform makes our models adaptable to new video sources as well.

To enable supervised modelling, we used the CVAT toolbox, which was discussed in [8]. The interface allowed us to quickly label bounding boxes and tracks across thousands of frames. The intuitive tools, like interpolation between keyframes, made labelling long footage feasible. Once the annotation was complete, we exported the labelled dataset into PASCAL VOC 1.1 for ML model training and traffic participant pattern analysis. CVAT.ai's cloud access and collaboration features made it optimal for this annotation task.

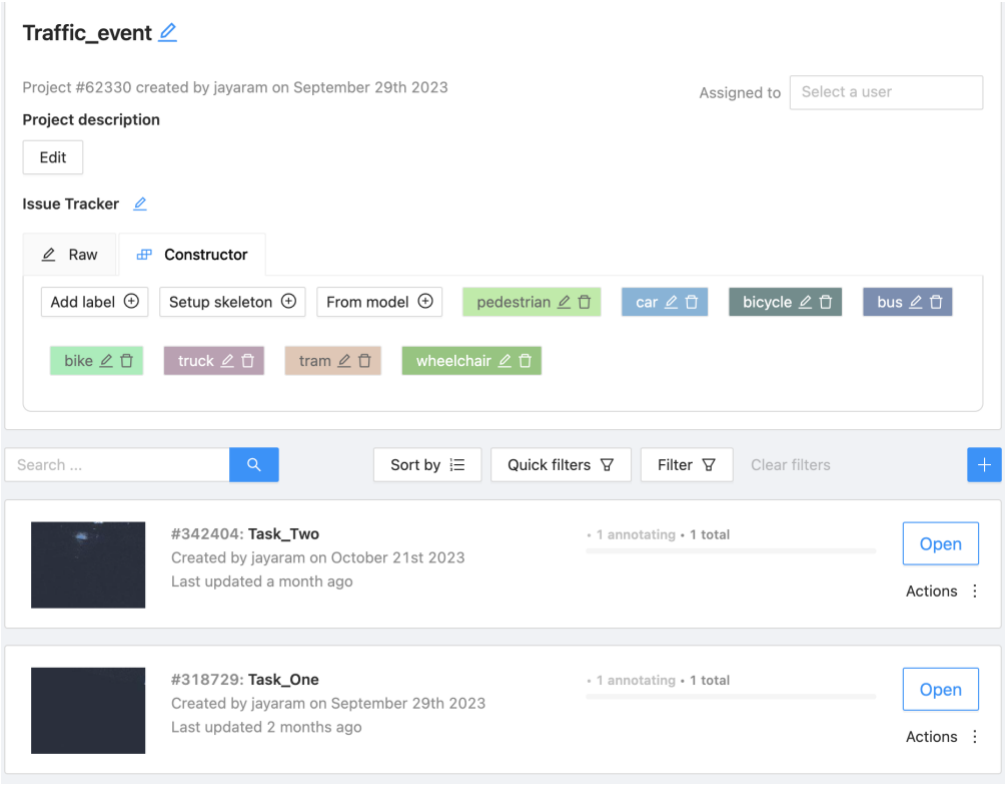


Figure 13: CVAT tool interface mentioning all the considered labels.

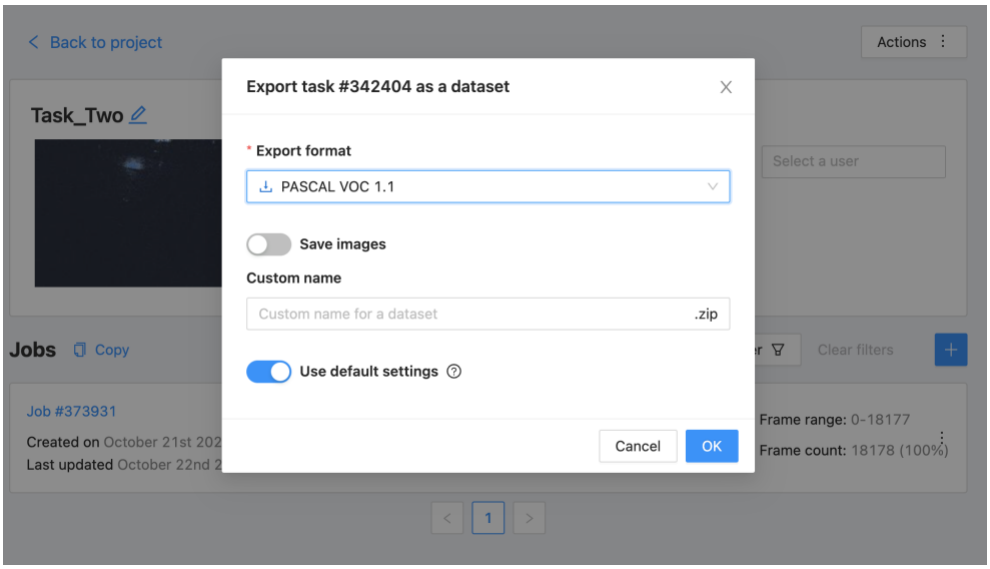
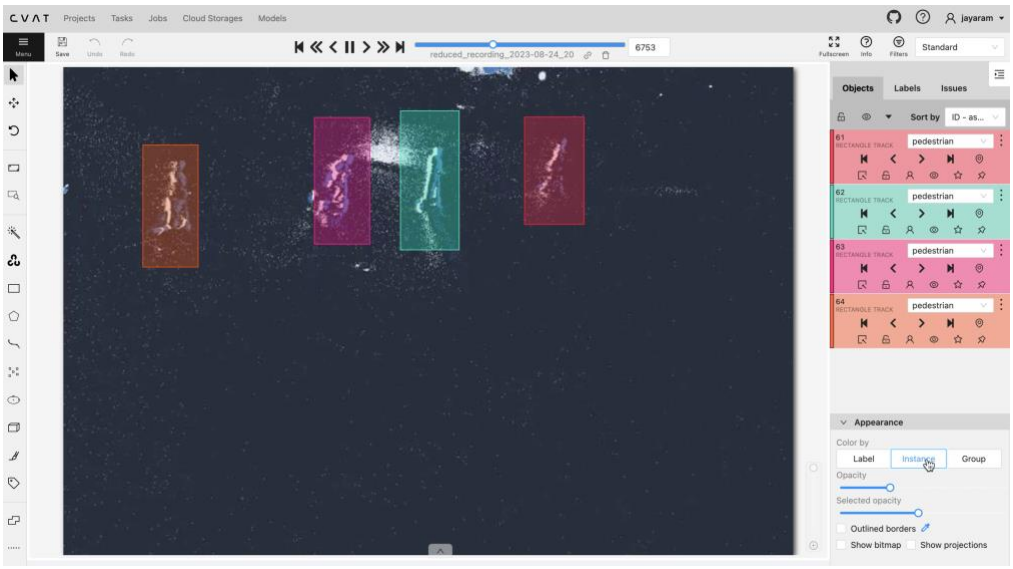
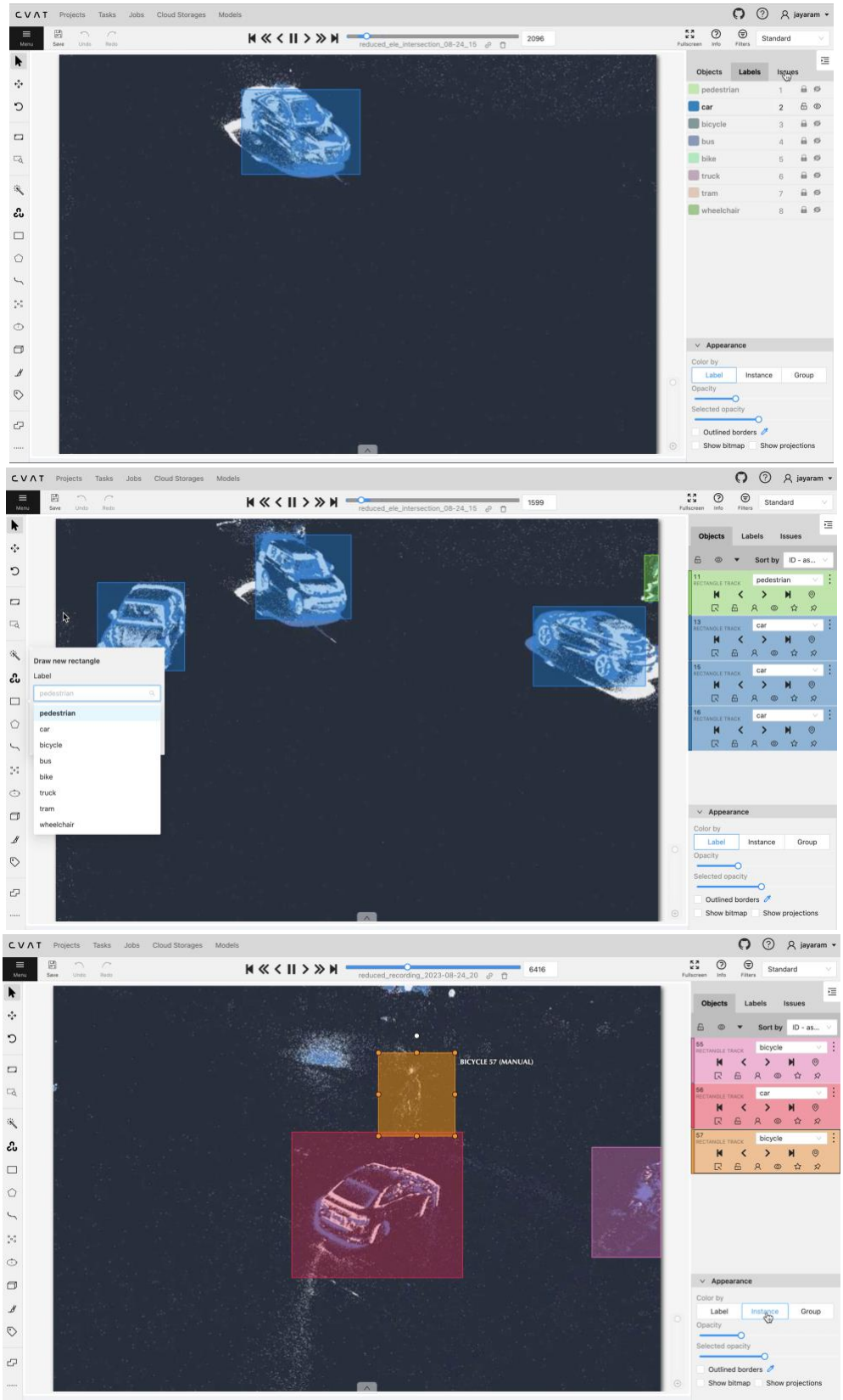


Figure 14: Exporting to PASCAL VOC and Task saving.

The foreground segmentation capability was specifically useful to separate dynamic and static pixels for encoding motion patterns. Annotation quality was improved by using interpolation across keyframes and consensus review across a distributed team. In total, I have annotated over 36000 frames for the event camera data collected at traffic intersections at ASU Tempe, at Brickyard, and various intersections for different daylight and weather conditions. We collected and annotated over 50 event data videos. The labelled dataset exported in PASCAL VOC format (Figure 14) can enable the training of various detection network architectures.

Figure 15 - Annotations labelled (For various intersection Participants, objects),All the images here feature various aspects in the tool along with the objects




CVAT
Projects
Tasks
Jobs
Cloud Storages
Models
jayaram
Objects
Labels
Issues
pedestrian
car
bicycle
bus
bike
truck
tram
wheelchair
Appearance
Color by
Label
Instance
Group
Opacity
Selected opacity
Outlined borders
Show bitmap
Show projections
Draw new rectangle
Label
pedestrian
car
bicycle
bus
bike
truck
tram
wheelchair
BICYCLE 57 (MANUAL)

2.5 Proposed Model for Supporting Model Development

After analyzing existing literature and consulting project sponsors and mentors, we proposed a multi-stage sensor fusion model that combines data from the event and RGB cameras. The purpose of this model is to leverage the strengths of these two modalities: the high temporal precision of events and the rich spatial information of RGB frames.

The first stage focuses on spatio-temporal feature extraction from both modalities through separate encoder pathways. The dynamic event data is accumulated into 4D event tensors that encode motion and activation over space and time. The RGB frames provide additional context for processing complex driving scenes.

In the second stage, the learned representations are fed into a Region Proposal Network (RPN) to generate 2D region candidates likely to contain foreground objects such as vehicles and pedestrians. This region generation benefits from precise motion cues from event features to localize objects, supplemented by global scene understanding from RGB encodings.

The third stage classifies the proposed regions into specific object categories and regresses their accurate 2D bounding box coordinates. This detection head fuses region-specific intermediate features from both pathways to leverage temporal event precision and RGB semantics.

Furthermore, the model proposes long-range temporal modelling modules for further analysis of behaviors over time. These modules include optical flow for pixel motion estimation, LSTM trackers for trajectory forecasts, and event accumulation units to identify anomaly patterns.

While this model is based on literature insights, it has yet to be implemented by the team. We plan to build this network, train it on our multi-modal data, and evaluate its performance on metrics such as intersection object detection accuracy. The representation learning framework is designed to synergistically combine the complementary strengths of the two sensor streams.

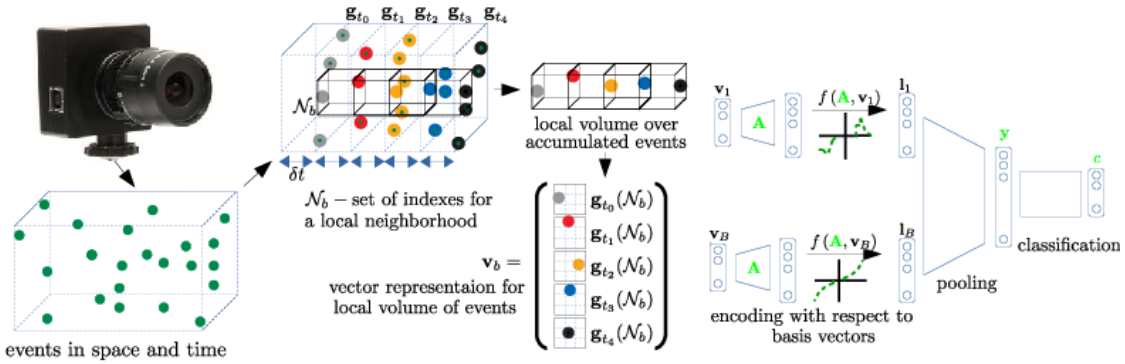


Figure 16: Model representation for event data flow

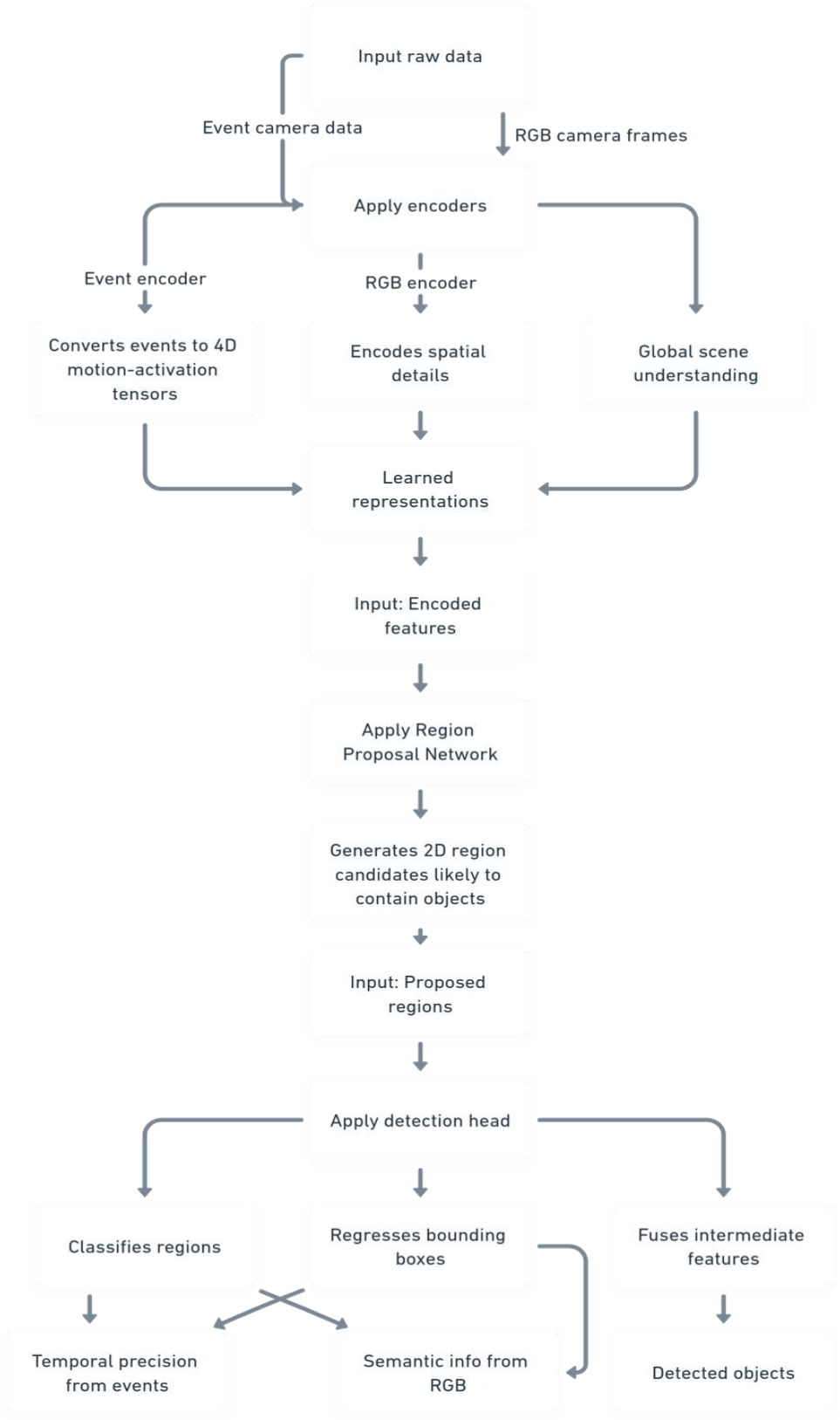


Figure 17: Model flow chart showing stages.

3. CONCLUSION

Throughout this project, we gained hands-on experience with event cameras, which can capture micro changes that are often missed by standard sensors. However, we encountered a few key challenges, such as synchronizing and formatting the unconventional event data for fusion with conventional RGB frames. After conducting a literature review and design exploration, we proposed an RGB + event model to leverage complementary motion and appearance information. The dataset and engineering foundation developed will empower the future execution of the Intersection Safety Challenge.

To summarize, this effort demonstrated the potential of asynchronous event cameras to better detect dynamic objects such as pedestrians and cyclists that are critical for intersection safety. Event cameras offer an analogue view of motion in the scene, as opposed to discrete sampled frames, which can miss important events in between captures. This makes them well-suited for capturing the continuous subtleties of slow-moving objects.

More broadly, this project highlighted the importance of selecting optimal sensors and fusing modalities to achieve robust perception systems. Although event cameras enable capturing temporal changes, RGB cameras still provide useful appearance information. Determining how to properly synchronize and encode the data for fusion is non-trivial but essential to unlocking the complementarity. Through the design process, the value of perceiving both space and time became increasingly clear.

From a personal perspective, working on applied perception systems imparted useful insights on trade-offs, data integration challenges, system cohesion principles, and end-to-end development thinking. For example, balancing performance metrics and handling the alignment of asynchronous data modalities proved difficult but instructive. Tackling the hardware and software facets holistically rather than sequentially was also an important learning experience.

The knowledge and dataset contributions made during this project will be invaluable for advancing towards safer autonomous vehicles. Most importantly, this project reiterated that challenges drive progress when approached with an open and persistent mindset. Pushing through the data hurdles imparted broader learnings that will echo through future perception endeavors.

3.1 Acknowledgements

I am grateful to Dr. Wishart for his invaluable CAV classes. His explanations have helped me to better understand the latest trends and relate them to my ongoing project. I would also like to extend my gratitude to Dr Wishart for giving me the opportunity to work with Dr. Yezhou Yang and Dr. Bharatesh Chakravarthi's team. During this period, I have learned more about cameras and perception. I look forward to continuing my work with the team, particularly in the areas of model development, testing, and deployment.

REFERENCES

[1] P. de Tournemire et al., “A Large Scale Event-Based Detection Dataset for Automotive,” CVPR Workshops, 2020.

[2] Guillermo Gallego and Davide Scaramuzza, “Events-based Vision: A Survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[3] Christian Brandli et al., “Real-Time, High-Speed Video Decompression using a Frame-and Event-based DAVIS Sensor,” IEEE International Symposium on Circuits and Systems (ISCAS), 2014.

[4] L. I. Triginer, “Deep Event-based Object Tracking and Multi-camera Fusion,” International Conference on Control, Automation and Diagnosis (ICCAD), 2022.

[5] Alex Zihao Zhu et al., “EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras,” Robotics: Science and Systems (RSS), 2018.

[6] P. de Tournemire et al., “A Large Scale Event-Based Detection Dataset for Automotive,” CVPR Workshops, 2020.

[7] Song et al., “How to calibrate your event camera: A pedagogical tutorial-style survey,” CVPR Workshops, 2021.

[8] H. Rebecq, D. Gehrig, D. Scaramuzza, “ESIM: an Open Event Camera Simulator,” Conference on Robot Learning (CoRL), 2018.

[9] Inikupilli et al., “Neuromorphic cameras,” Nature Reviews Physics, 2021.

[10] Barua et al., “Direct face detection and video reconstruction from event cameras,” Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV), 2021.

[11] Amir et al., “A Low Power, Fully Event-Based Gesture Recognition System,” CVPR, 2017.

[12] Aung et al., “Event-based detection, tracking and surveillance: concepts, challenges and opportunities,” Nature Machine Intelligence, 2021.

[13] Chen et al., “Multi-modal Sensor Fusion for Joint 3D Estimation and Scene Understanding via Sequential Predictions,” IEEE Trans. Intelligent Vehicles.

[14] Mayer et al., “A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation,” CVPR, 2016.

[15] Xue et al., “FUSEMODNET: Real-time Camera and LiDAR Based Moving Object Detection for Robust Low-light Autonomous Driving,” ICCV Workshops, 2019.

[16] Lianos et al., “VSO: Visual Semantic Odometry,” ECCV, 2018.

[17] Gallego et al., “A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation,” CVPR, 2018.

[18] CVAT, “Computer Vision Annotation Tool,” https://cvat.ai

[19] Intersection Safety Challenge Website, https://its.dot.gov/isc/

[20] RFI Summary Report, FHWA-JPO-23-986, February 2023.

[21] USDOT Intersection Challenge Submission Page, https://www.challenge.gov/?challenge=us-dot-intersection-safety-challenge

[22] Event-Based Concepts, https://docs.prophesee.ai/stable/concepts.html

[23] Prophesee Event Camera EVK4, https://www.prophesee.ai/event-camera-evk4/

[24] Synchronization Manual, https://docs.prophesee.ai/stable/hw/manuals/synchronization.html

[25] Event Camera Calibration Paper, https://tubrip.github.io/eventvision2021/papers/2021CVPRW_How_to_Calibrate_Your_Event_Camera.pdf

[26] Sensor Fusion Resources, https://paperswithcode.com/task/sensor-fusion

[27] ASU SCAI,
https://scai.engineering.asu.edu/research-labs/

[28] Event-based Vision Research,
https://rpg.ifi.uzh.ch/research_dvs.html