University Nice Côte d'Azur
MSc2 Data Science and AI

and

Computer Science, Signals and Systems Laboratory
of Sophia Antipolis (i3S)

Attention Model for Event Data

Intern Student
Huyen Trang NGUYEN

Internship Supervisors
Jean MARTINET
Michel RIVEILL

Sophia Antipolis, 20 August 2024

Abstract.

This internship project delves into cognitive attention models (both bottom-up and top-down) designed for event data, with a focus on applications in autonomous driving. Traditional computer vision models primarily use frame-based images, which generate excessive redundant data and require substantial computational power. In contrast, event cameras—or dynamic vision sensors (DVS)—present an innovative solution by recording only the changes in brightness at each pixel, thereby minimizing data redundancy and reducing power consumption. Although event data is potential in both research and industry, there exists only few attention models tailored for this data type, especially top-down attention model. The goals of this internship include organizing and synthesizing current knowledge on event data and cognitive attention models, and applying these insights to practical detection tasks using the DSEC-Detection dataset. The first contribution resulted in a comprehensive survey that offers significant guidance for future studies in this field. The next contribution is adapting YOLO (bottom-up) and YOLO-World (top-down) models to event data, marking a novel attempt to fine-tune these models for dynamic vision sensors. The results of this internship are promising, suggesting that with further development and increased computational resources, these models could significantly enhance the performance of vision systems for the usage of event data, particularly in autonomous driving.

Keywords: Cognitive attention · Top-down and bottom-up · Event data.

Table of Contents

1 Introduction

1.1 Introduction about i3S

Established in 1989, the i3S Laboratory (Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis) is a joint research unit (UMR 7271) affiliated with CNRS and Université Côte d'Azur (UCA), and collaborates with Inria on five joint projects. i3S, one of UCA's largest research units, is located in Sophia Antipolis, a major European technopole. This insititute comprises around 300 members, including 80 faculty from UCA's various departments, 20 CNRS researchers, 10 Inria researchers, and about 100 PhD students, 25 postdocs, and 60 interns. The laboratory's research spans Computer Science, Electrical Engineering, Signal and Image Processing, Control Systems, and Robotics, aligning with sections 27 and 61 of CNU and sections 6 and 7 of CoNRS. I3S is organized into four divisions: COMRED, MDSC, SIS, and SPARKS, and collaborates with Inria on projects like Coati, Kairos, Maasai, Morpheme, and Wimmics.

1.2 Introduction about internship project

1.2.1 The context Modern computer vision systems often rely on frame-based images, which generate huge amounts of redundant data and demand significant computational resources. In contrast, event data has emerged as a potential field in both research and industry due to its energy efficiency, high time resolution, and other benefits. Besides, directing attention to the most informative elements of each image, while disregarding less relevant details, is critical in computer vision for reducing computational load, improving efficiency, and ensuring that models focus on the most relevant aspects of the visual data. This internship centers on studying attention model for event data for detection tasks in autonomous driving application.

Event data In the field of computer vision, the quest for low-latency, energy-efficient vision systems has catalyzed the development of innovative models capable of processing vast amounts of visual data with high efficiency. Traditional frame-based cameras, despite their effectiveness, produce a fixed amount of data per frame, leading to redundancy and significant processing demands. Event cameras, also known as dynamic vision sensors (DVS), present an alternative by capturing only changes in brightness at each pixel, thereby reducing data redundancy and power consumption. These cameras offer high temporal resolution and low latency, making them particularly suitable for dynamic and real-time processing environments. Figure 1 [2] shows an example of the difference between data captured by a frame-based camera (left) and an event camera (right) in low-light conditions. The event camera image provides more details, such as the person walking (in the lower right of the image) and distant background elements like buildings, trees and cars, which are less visible or absent in the frame-based camera image. The detailed characteristics of event cameras and event data will be explored further in the following section.

Fig. 1: Example of data captured by frame-based camera (left) and event camera (right). Source: [2]

Attention in computer vision *It it is crucial to specify that the term* ***"attention"*** *in this internship refers to* ***"cognitive attention"***, which is different from the attention mechanisms in Transformer architecture. Both cognitive attention and attention mechanisms in Transformer will be clarified further in the Section 2.2.

The concept of attention (or cognitive attention), inspired by human visual system, refers to how the brain selects and analyses important information according to specific tasks. Attention mechanisms can be categorized into bottom-up attention, driven by the inherent saliency of sensory input, and top-down attention, guided by higher-level cognitive factors such as task relevance and prior knowledge. Bottom-up attention focuses on the inherent features of stimuli, while top-down attention integrates contextual information and cognitive biases to direct focus based on the task at hand. The potential applications of these integrated attention mechanisms are vast, including robotics, autonomous driving, gesture recognition, and surveillance. Reducing computational load and improving response times, these systems can operate more efficiently in real-time scenarios, offering significant advancements in performance and energy efficiency.

1.2.2 The internship objectives

Despite the rapid growth and promise of event data in research and industry, there is only few existing works for attention models for this type of data. This internship is part of an interdisciplinary collaborative research project, *Neuromorphic Attention Models for Event Data* (NAMED[1]), that aims to develop computational models of cognitive attention that combine both event sensors (simulating peripheral vision) and RGB sensors (simulating central vision), with improved power efficiency in embedded electronics, mimicking the hu-

[1] Project number ANR-23-CE45-0025-02, URL: https://wwww.i3s.unice.fr/named/

man visual system to selectively focus on regions of interest. The objectives of this internship is a pioneering attempt to explore the potential of attention model within the realm of event data, both theoretically and in practical applications, particularly for object detection tasks for autonomous driving applications.

1.2.3 The main contributions of the internship Despite the growing interest in attention models for event data, a comprehensive understanding and well-structured terminology for this domain are still lacking in existing studies. The primary contribution of this internship is to address this gap by organizing and structuring the knowledge surrounding event data and attention models applied to event data. This involves a detailed analysis and summary of studies related to event data, existing datasets classified by task categories, and attention models, presented in a comprehensive survey. This survey not only provides valuable insights for future research but also serves as a guide for further exploration in the field. Additionally, the survey is currently accepted for a poster presentation for the ICONIP 2024 Conference [2] and is presented in Section 2 (Literature review).

The next main contribution is the application of YOLO [29] and YOLO-World [8] to the DSEC-Detection dataset [1] – a large and recent event dataset used for detection tasks in autonomous driving scenarios. YOLO (You Only Look Once) is an object detection algorithm that revolutionizes the field by detecting objects in images or videos with remarkable speed and accuracy, treating object detection as a single regression problem rather than a series of classification and localization tasks. YOLO-World is an open-vocabulary detection model that allows for the definition of specific classes prior to detection. YOLO is chosen as the bottom-up model due to its ability to learn and adapt from data without predefined guidance, while YOLO-World is selected as the top-down model because it allows for task-specific focus by enabling predefined object detection through class specification before training. The results from applying YOLO and YOLO-World to DSEC-Detection dataset are experimental tests of bottom-up and top-down attention models in event data respectively. These results are promising and suggest strong potential for future research and practical applications.

1.2.4 Structure of the internship report The structure of this report is organized as follows: Section 1 provides an introduction to the internship, outlining its key objectives and contributions. This is followed by an overview of event cameras, event data, and existing datasets in Section 2. In Section 2.2, we delve into the concepts of visual attention in computer vision, exploring regions of interest, eye gaze, and the interaction between bottom-up and top-down mechanisms for both RGB and event data. Section 4 focuses on the DSEC-Detection dataset and the data preprocessing steps undertaken during the internship. The report then presents a detailed discussion of the YOLO and YOLO-World models, including their fine-tuning and

[2] ICONIP 2024, URL: https://iconip2024.org/

evaluation processes in Section 3. In Section 5, we analyze the results and provide further insights. Finally, the report concludes in Section 6 with recommendations for future research directions.

2 Literature review

In this section, we will explore the fundamental principles of event cameras, event data, existing methodologies working on event data and attention for event data in computer vision.

2.1 Event cameras and event data

We start by examining the concept of event cameras, event data, and existing event datasets. Additionally, we discuss the methodologies used for processing event data, highlighting both synchronous and asynchronous approaches. This overview sets the stage for a deeper examination of attention models in event-based vision systems.

2.1.1 Event camera , also known as dynamic vision sensors (DVS), represents a significant shift from traditional frame-based camera. Unlike frame-based camera that captures image frames at fixed intervals, event camera only records changes in brightness at each pixel, independently and asynchronously [51]. Each change in light is considered an "event", which can be Positive or Negative (On or Off), corresponding to increases or decreases in brightness. Figure 2 [51] shows the difference between data captured by a frame-based camera (left) and an event camera (right). On the left is a screenshot captured from a video recording, depicting an arm waving in front of the camera. Conversely, the image on the right is a frame generated from event data, where only the arm is visible due to its motion causing changes in light. In this event frame, other objects such as the table and whiteboard are not captured by the event camera because they remain stable, resulting in no changes in light on these objects, unlike in the frame image on the left.

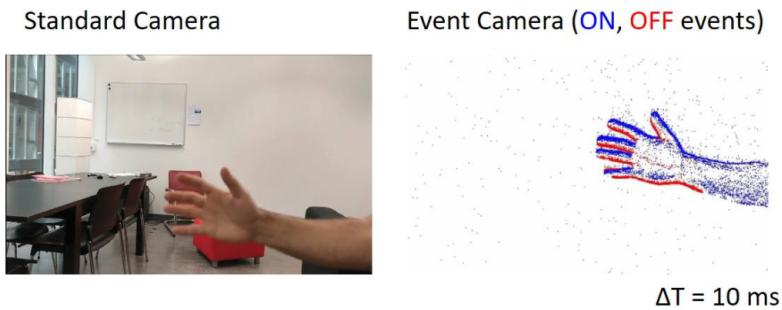

Fig. 2: Difference between data captured by frame-based camera (left) and event camera (right). Source: [51]

Key Features of Event Cameras:

Asynchronous Data Capture: Each pixel in an event camera operates independently, detecting changes in light intensity and outputting data whenever a change exceeds a certain threshold. This means that pixels only generate data when there is movement or lighting change, significantly reducing redundancy.

High Temporal Resolution and Low Latency: Because they do not operate at a fixed frame rate but rather respond instantly to changes, event cameras can have extremely high temporal resolution, often in the order of microseconds. This allows them to capture very fast-moving objects effectively, significantly reducing motion blur that is commonly seen in traditional video when capturing rapid movements. The low latency in processing these changes further enhances their capability to deliver sharp images in dynamic environments.

Low Power Consumption: Since data is only generated by pixels when changes are detected, event cameras typically consume less power compared to traditional cameras, which continuously capture frames at a regular interval, regardless of whether the scene has changed.

High Dynamic Range: Event cameras excel in environments with high dynamic range (HDR) situations, such as scenes that quickly transition between very dark and very bright conditions. Traditional cameras often struggle with such extremes due to fixed exposure settings.

2.1.2 Event data , captured by event camera, is typically represented in the form of (x, y, t, p), where x and y denote coordinates, t represents the time stamp when light changes (event) occur, and p indicates polarity, which can be either positive or negative, corresponding to lighter or darker changes in brightness, respectively. This results in a stream of events rather than frame-based cameras, capturing only the dynamic parts of a scene.

Key Features of Event Data:

Sparse Output: The data from event cameras is inherently sparse and consists only of the changes in the scene, not the static parts. This makes it highly efficient but also different in nature from typical video data, requiring different processing algorithms.

Time Stamped Events: Each event is time-stamped, providing precise information about when the change occurred, which is crucial for applications requiring accurate temporal measurements.

Data Stream: The output is a continuous stream of events, which can be processed in real-time or stored for later analysis. The stream-based nature of the data is well-suited for real-time systems such as robotics and autonomous driving.

In summary, event cameras and Event data offer unique advantages for dynamic and real-time processing environments, standing out particularly where traditional cameras fail due to limitations in dynamic range, speed, and power efficiency. These

capabilities make them a powerful tool in numerous technology applications such as robotics, automotive, surveillance, etc.

2.1.3 Existing event datasets , which have emerged in recent years, vary widely in structure and recording conditions, based on their specific needs and applications.

Event datasets for object recognition and 3D object perception include the Caltech-256 Dataset [25], which captures objects in motion with event cameras, and CIFAR10-DVS [33] and N-Caltech101 [45] which adapt CIFAR10 and Caltech101 images respectively for event cameras in recognizing categories. The Combined Dynamic Vision/RGB-D Dataset [61] provides scenarios for object recognition using synthetic, color, and depth sensors, while MVSEC [65] [66] and Event Camera Motion Segmentation Dataset offer extensive 3D object recognition and segmentation for various vehicles and objects indoor and outdoor. The N-CARS dataset supports specifically for cars classification [52]. Event Moving Object Detection and Tracking [43] includes event data for tracking moving objects. Slow-motion card symbol recognition is analysed by SLOW-POKER-DVS dataset. Besides object recognition, some event datasets supporting gesture recognition include the DHP19 [6] which presents a human pose dataset in event data. The DVSMOTION20 [3], ROSHAMBO17 [39] and SL-ANIMALS-DVS [58] [57] explore for human sign languages or hand moving.

For autonomous driving, DDD20 [24] and DSEC [19] [20] datasets offer extensive data on road scenes and vehicle movements in various brightness scenarios. The DET: A High-resolution DVS Dataset for Lane Extraction [9] focuses on lane extraction with labeled images. DND21: DeNoising Dynamic Vision Sensors Dataset [23] address action recognition and denoising for surveillance and driving use cases. Driving videos are further explored in the Driving Event Camera Dataset [49], including high speed driving datasets, and DVS09 [12] [36], which is utilized for simple outdoor driving in day-light condition. In addition, EDFLOW21 [38] focuses on driven-flow with event data. Pedestrian actions event data is analysed in Neuromorphic Vision Dataset for Pedestrian Detection, Action Recognition, and Fall Detection [42] and GEN1 Automotive Detection Dataset [11].

In conclusion, existing event datasets offer a diverse range of applications, from object and gesture recognition to autonomous driving. These datasets are tailored to specific tasks and recording conditions, providing valuable resources for advancing research in event-based vision systems. The growing availability and variety of these datasets highlight the increasing importance and potential of event data in addressing complex real-world challenges in computer vision. In the following part of the report, we will analyse the methods of computer vision for event data.

2.1.4 Existing research works on event data for computer vision tasks Despite the numerous advantages of event cameras, such as high resolution, low latency, and reduced motion blur, there are notable challenges in utilizing this type of data for object detection, particularly because most existing models are optimized

for frame-based data. To address Event data, two primary approaches are typically employed as following.

Synchronous Methods

These involve using Artificial Neural Networks (ANNs) or Convolutional Neural Networks (CNNs) that are designed for standard images. This approach often includes converting Event data into frame-like formats to leverage existing computer vision models. Techniques for this transformation include: Creating event frames by aggregating positive and negative events into separate channels, though this method tends to discard temporal information [40]. Representing events in a 3D voxel grid, where each voxel aggregates the sum of ON and OFF events over time, preserving some temporal information but losing polarity details [67] [66]. Developing a 4D Event Spike Tensor representation (x, y, t, p) that maintains both temporal dynamics and polarity information, enhancing the data's richness for further processing. Nonetheless, this method introduces processing complexity [17].

After transforming Event data into image frames, the adapted data can be processed using existing models that are optimized for frame-based data. This data transformation allows the application of computer vision techniques originally designed for traditional video formats. As a result, various of works such as video reconstruction [49], object detection [49] [18], etc. can be effectively undertaken.

Asynchronous Methods

Spiking Neural Networks (SNNs) are exceptionally well-suited for processing Event data from event cameras, offering significant advantages due to their ability to handle asynchronous, temporal events through discrete, time-sensitive spikes [5]. These networks are highly energy-efficient, activating only in response to specific sparse stimuli, which aligns perfectly with the sparse outputs from event cameras that signal only during changes in light intensity. SNNs also mimic biological neural processing by firing only when a predefined threshold is reached, efficiently managing polarity information to enhance detection and classification of changes in the visual scene [22].

Each method has its own set of research works exploring its potential. The synchronous approach often involves transforming event data into more familiar formats for traditional neural networks, whereas the asynchronous method focuses on directly harnessing the unique properties of event data through SNNs. Both strategies are critical in pushing the boundaries of what can be achieved with Event vision systems.

In summary, this section provides an in-depth overview of event cameras and event data, highlighting their unique characteristics compared to traditional frame-based cameras. We also discusses various event datasets, categorized by their target applications, including object recognition, gesture recognition, and autonomous driving. Additionally, we explores synchronous and asynchronous methods for processing event data, emphasizing the potential of these technologies in advancing research in dynamic, real-time computer vision systems. In the next section, we ex-

plore attention mechanisms for computer vision, focusing on both RGB and event data.

2.2 Attention for event data in computer vision

For the purpose of this internship, it is crucial to clarify that the term ***"attention"*** used throughout refers specifically to ***"cognitive attention"*** rather than the ***"attention mechanisms"*** associated with Transformers. In this subsection, we delve into the concept of attention as it applies to event data in computer vision, distinguishing between cognitive attention and the attention mechanisms commonly used in Transformer architectures. We will explore both bottom-up and top-down attention models, which serve as fundamental mechanisms in directing focus within a visual scene. While bottom-up attention is driven by the intrinsic saliency of stimuli, top-down attention is guided by task-specific goals and expectations. This section aims to provide a comprehensive analysis of these models, particularly in the context of event data, where research is still in its early stages. We will analyse existing models to those used for both RGB data and event data and highlight their potential for future developments in computer vision applications.

2.2.1 Cognitive attention in computer vision We highlight below the distinction between cognitive attention and attention mechanism of Transformer architecture for computer vision in general.

The attention mechanism [59] in the Transformer architecture operates originally in the domain of Natural Language Processing (NLP). At its core, the Transformer attention mechanism enables the model to capture long-range dependencies within sequences by weighing different parts of the input sequence differently, thereby focusing more on relevant information and disregarding noise. Additionally, recent advancements, such as the Vision Transformer (ViT) [13], have extended and adapted the attention mechanism from Transformer to computer vision by segmenting images into fixed-size patches, embedding them, and employing a Transformer encoder. This technique enhances image classification with attention-based processing.

Unlike attention for Transformers, cognitive attention involves higher-order cognitive processes beyond sensory perception, such as memory, decision-making, and problem-solving [56]. Cognitive attention mechanisms enables the brain to prioritize and efficiently process sensory information from the environment. It can manifest across various sensory modalities such as auditory and tactile perception. Tsotsos and colleagues define attention as "the process by which the brain controls and tunes information processing." [54]. The human visual system only allows for high-resolution visual information to be encoded from the fovea (the central 2° of vision). Visual quality falls off rapidly and continuously from the center of gaze into a low-resolution visual surround. As a result, we constantly move our eyes (saccade) to redirect the fovea towards a new area where the visual information will be acquired

when the eye is stable (fixation). Thus, due to the structure of our visual system, human vision depends on eye movements. Understanding the factors that guide these eye movements is therefore an important component of understanding how humans process visual information and has a wide range of applications in computer vision. A two-component framework for attentional deployment has emerged in the early 2000', suggesting that the observer selectively directs attention to objects in a scene using both bottom-up, image-based cues, and top-down, task-dependent cues [27], as described below.

Visual attention is attracted by salient stimuli that *pop out* from their surroundings. Bottom-up, or stimulus-driven, selection is said to occur when attention captured by properties of the stimulus even if they are irrelevant to the current task. Some stimuli are intrinsically conspicuous or salient and spontaneously and involuntarily attract attention. Saliency, which is independent of the nature of the particular task, operates very rapidly. This suggests that saliency is computed in a pre-attentive manner across the entire visual field.

But attention can also be voluntarily directed to objects of current importance to the observer. Top-down, or goal-directed, selection is said to occur when the observer's knowledge or beliefs about the task determine what is selected in the visual field. Attention adapts the visual system to its dynamic needs. For example, when individuals are highly engaged in a problem-solving task (top down) or when they have to detect targets rapidly on a screen (bottom up), visual inputs are processed differently. When attempting to construct a general model of visual attention, one has to take into account the fact that these different strategies are more or less induced by the tasks performed by the participant.

In the real world, we constantly move our eyes to direct the high-resolution fovea towards points of interest in the environment. This situation is different from in-laboratory research tasks, where stimuli are displayed on a screen and the useful visual field is smaller. Such studies involve foveal rather than peripheral vision. The scientific literature is mainly focused on modelling bottom-up rather than top-down attention. But above all, it is much less common to find models that take both sources of information into account in a dynamic way [62].

In the following sub-sections, we delve into key concepts in attention for computer vision, such as top-down and bottom-up attention, Region of Interest, and eye gaze.

2.2.2 Bottom-up and top-down attentions Bottom-up attention, often referred to as exogenous attention, is driven by sensory input and operates rapidly and involuntarily [31] [16]. One prominent concept in bottom-up attention is *saliency*, which denotes the degree to which a stimulus stands out from its surroundings [48]. Saliency-driven attention is guided by the conspicuousness of visual features such as color, orientation, and motion, with salient stimuli attracting attention spontaneously. Computational models, such as saliency maps, have been developed to simulate bottom-up attention, wherein saliency values are computed across the vi-

sual field, facilitating the detection of regions likely to capture attention [60]. These models provide insights into how visual stimuli compete for attentional resources based on their saliency, offering a framework for understanding early stages of visual processing.

In contrast to bottom-up attention, top-down attention is endogenous, driven by internal cognitive factors such as goals, expectations, and task relevance. This form of attention is characterized by its voluntary nature and its ability to prioritize specific stimuli based on cognitive goals and expectations [31] [16] [37]. Top-down attention can modulate sensory processing, biasing attention toward stimuli relevant to the observer's goals while filtering out irrelevant information. While bottom-up attention is primarily concerned with the saliency of sensory input, top-down attention operates at a higher cognitive level, integrating contextual information and cognitive biases to guide attentional selection [44] [10]. Modeling top-down attention poses challenges due to its complexity and its dependence on internal cognitive states, making it a subject of ongoing research in cognitive neuroscience and computational modeling.

2.2.3 Region of interest and eye gaze Regions of Interests (RoI) and scan paths (sequence of points of gaze) provide complementary insights into attentional mechanisms and cognitive processing. The gaze is focused on a point on the scene or screen and RoI represent the zone of what is actually seen in a single glance in central vision. RoI selections, driven mainly by top-down attention and task-specific goals, offer a direct insight into cognitive processing and perceived interest in visual scenes [14]. In contrast, scanpaths reflect the interplay between bottom-up and top-down attention, providing temporal information about the observer's visual behavior [55]. While RoI selections provide spatially defined regions of interest based on cognitive factors, scanpaths offer continuous temporal information about the observer's attentional allocation during scene processing [14]. Despite differences in spatial coverage and temporal resolution, both approaches contribute to understanding attentional mechanisms and cognitive processes involved in visual perception.

In the subsequent sections of this report, we will delve deeply into the analysis of both bottom-up and top-down attention models, focusing on their application to RGB data as well as event data. This thorough examination is designed to provide a comprehensive understanding of the current research landscape in this field, making it a valuable resource for guiding future studies. While attention models for RGB data have been extensively studied, the primary focus of this internship is on event data, where particular emphasis will be placed on models specifically designed for this emerging data type. It is important to highlight that top-down attention models for event data are still in the early stages of exploration, representing a novel and promising area for further research. Table 1 offers a summarized overview of the bottom-up and top-down attention models that will be discussed in detail throughout the report, setting the stage for a deeper exploration of these concepts.

Table 1: Summary of Top-down and Bottom-up Attention Models for RGB and Event Data

	RGB Data	Event Data
Bottom-up	Itti et al. (1998) [28] Martinet et al. (2009) [41] Wang et al. (2010) [60] Gregor et al. (2015) [21]	Li et al. (2019) [34] Iacono et al. (2019) [26] Yao et al. (2021) [63] Liang et al. (2021) [35] Sun et al. (2022) [53] Gruel et al. (2022) [22] Bulzomi et al. (2023) [5]
Top-down	Zheng et al. (2015) [64] Radford et al. (2021) [47]	Cancini et al. (2019) [7] Kong et al. (2024) [30]

2.2.4 Top-down and bottom-up attention for RGB data In the domain of computer vision, much research has explored attention approaches for analyzing RGB data. In this section, we will briefly review top-down and bottom-up models for RGB data before shifting more deeply into those for event data in the section 2.2.5.

In 1998, Itti, Koch, and Niebur [28] introduce a computational model of visual attention that mimics the early stages of visual processing in primates. The model creates a saliency map by integrating various visual features such as intensity, color, and orientation at different spatial scales. These features are merged to emphasize the most visually prominent areas within a scene. The saliency map is then used to direct attention towards these regions, facilitating rapid analysis of the scene. The primary goal of this model is to replicate the way humans and primates quickly identify significant areas in their visual field without relying on prior knowledge or specific tasks. This model is classified as a bottom-up approach because it depends entirely on the characteristics of the visual stimuli to decide where to focus attention. In Itti et al.'s model, attention is given to areas that stand out due to distinct visual features, such as high contrast or unique colors, rather than any pre-determined expectations or goals. This approach is particularly effective for initial scene analysis, where quickly identifying salient features is essential.

The study by Martinet et al. (2009) [41] explores gaze analysis as a tool for evaluating the quality of visual media, focusing on a bottom-up visual attention approach. This method utilizes viewers' eye movements to determine how effectively visual media communicates its intended message. By tracking and analyzing where and how long viewers look at different parts of an image or video, the research identifies the elements that naturally attract attention. This approach depends on the visual characteristics of the media itself, such as color, texture, and movement, emphasizing a data-driven method that highlights the most eye-catching features and thus representing a bottom-up perspective in visual attention analysis.

Wang et al. (2010) [60] examines the link between bottom-up (or visual salience) and top-down (or visual importance) by two experiments. The first experiment in-

volved participants rating the importance of hand-segmented objects in images. The second used eye-tracking to determine visual saliency based on where participants naturally looked in images without a task. The findings reveal a moderate correlation between importance maps and saliency maps. Saliency maps highlight shape and color, with early focus on human and animal faces. In contrast, importance maps prioritize categories such as human and animal faces, influenced by artistic meaning of the image. It also revealed a strong relationship between visual salience and visual importance in the first two seconds of viewing, indicating that bottom-up attention efficiently identifies primary subjects initially. This research highlights the complementary roles of bottom-up and top-down attention mechanisms.

The Deep Recurrent Attentive Writer (DRAW) network by Gregor et al. (2015)[21] for image generation features a unique spatial attention mechanism inspired by human eye foveation. The architecture of DRAW model contains encoder and decoder, both are recurrent networks, which iteratively build images through sequential modifications. The network uses dynamically updated attention, focusing selectively on input and output regions with 2D Gaussian filters to create image patches with varying locations and zoom levels. This bottom-up attention allows DRAW to adapt to different object scales and positions based on database, enhancing its performance in complex visual tasks.

Zheng et al. (2015) [64] developed Fixation NADE model, which uses visual attention by directing the recognition process through a sequence of task-specific fixations. Unlike traditional models that densely extract features over an entire image, Fixation NADE simulates human visual attention by learning both what features to extract and where to extract them, using a fixation policy. This approach combines both bottom-up and top-down attention mechanisms, where bottom-up attention is influenced by the data observed at each fixation point, and top-down attention is guided by task-specific objectives (gender classification and expression classification). Key contributions of this model include its autoregressive architecture for fixation-based recognition and its improved performance compared to earlier models like Fixation RBM [15].

CLIP (Contrastive Language-Image Pre-Training) model by Radford et al. (2021) [47] aims to develop visual representations using natural language supervision and employs a contrastive learning method to align images with their corresponding text descriptions. The goal of CLIP is to harness the rich information in natural language to enable zero-shot transfer learning, allowing the model to adapt to a variety of tasks without additional labeled data. The architecture includes an image encoder and a text encoder that work together to identify the correct image-text pairs. CLIP excels in multiple computer vision benchmarks, demonstrating strong performance in tasks such as action recognition, image description, or fine-grained object classification etc. The model leverages top-down attention by using textual descriptions to guide the focus on relevant image features. This innovative approach

highlights the potential of natural language supervision in developing versatile and transferable visual models.

2.2.5 Top-down and bottom-up attention for event data In contrast to RGB data, research of attention focusing on event data has been less prevalent, primarily due to its unique characteristics that demand specialized analysis. While there exists some studies of bottom-up, it is noteworthy that research on top-down models for event data remains relatively novel, with only a limited number of existing work. This could be due to the complexity of top-down approaches where the model depends significantly on specific predefined-tasks. However, top-down models remains a promising avenue for future research.

Li's 2019 paper [34] combines event data with frame-based images to enhance object detection for autonomous vehicles. The model employs two distinct processing streams: one utilizes convolutional neural networks (CNNs) for frame-based data, and the other leverages spiking neural networks (SNNs) for event data, which produce visual attention maps. By combining the high temporal resolution and dynamic range of event cameras with the detailed spatial information from traditional cameras, this approach aims to improve detection accuracy, particularly in scenarios with rapid motion and varying lighting conditions. The fusion of outputs from both streams is managed using Dempster-Shafer theory [32], resulting in improved performance as demonstrated on the DDD17 dataset [4]. The event vision system in this paper operates using a bottom-up approach. This method captures data only when there are changes in the visual field, which naturally highlights regions with motion and activity. This attention mechanism enhances the model's ability to detect vehicles effectively in challenging environments, highlighting the benefits of combining neuromorphic and conventional vision technologies in autonomous driving.

Iacono et al. (2019) [26] focuses on modifying a proto-object attention model to function with neuromorphic event cameras, aiming to enhance the iCub humanoid robot's visual processing capabilities. Utilizing a bottom-up attention mechanism, the model processes the event data through three layers: center-surround filtering, border ownership cells, and grouping cells. These layers decompose the visual scene into proto-objects, with saliency determined by changes in contrast and edges. The approach is validated with both static and moving objects, showing that higher object speeds lead to increased event counts, which enhances saliency. This model offers an efficient, low-latency method for robotic vision in dynamic settings, ensuring that the robot can effectively focus on relevant stimuli in its environment.

Yao et al.'s 2021 research [63] presents an innovative method for processing spatio-temporal event streams using Spiking Neural Networks (SNNs). The main goal of the study is to improve the accuracy and efficiency of SNNs by incorporating a temporal-wise attention mechanism. This model, named TA-SNN (Temporal-wise Attention Spiking Neural Networks), selectively filters out non-essential event streams (or denoising) during the inference phase based on their importance, which

is determined during the training phase. This optimizes computational resource usage. The model operates primarily with a bottom-up attention mechanism, as it processes event data by dynamically focusing on significant temporal segments determined by the data itself. The TA-SNN model achieved state-of-the-art results in various classification tasks, such as gesture recognition, image classification, and spoken digit recognition, demonstrating its effectiveness in handling sparse and uneven event streams.

In their paper in 2021, Liang et al. [35] introduces an innovative approach for Event object detection that employs a lightweight spatial attention mechanism. The primary goal is to improve the accuracy and efficiency of detecting objects in event data, especially by minimizing noise and enhancing multi-scale feature maps through the integration of shallow features. The core method involves encoding the event data into maps using techniques like Surface of Active Events (SAE) and Histogram of Intensities (HIS), followed by the application of a Canny edge detector to highlight relevant features. A key innovation is the Spatial Attention Module (SAM), which integrates multi-scale spatial features with the object detection framework. This module focuses on areas with significant event activity, enhancing the detection of moving objects by concentrating computational resources on RoI. The model utilizes a bottom-up approach, starting with raw event data and applying spatial attention to highlight significant features. By leveraging this lightweight attention mechanism, the system effectively filters out noise and improves detection accuracy, ensuring better performance in real-time automotive scenarios where rapid response and accuracy are crucial.

In the realm of autonomous driving, attention using event data is crucial for improving the system's ability to handle dynamic environments and ensure safety. Sun's 2022 research [53] introduces the Event Fusion Network (EFNet) designed to mitigate motion blur in images captured by conventional frame-based cameras. EFNet employs a symmetric cumulative event representation and integrates an Event-Image Cross-modal Attention (EICA) module to effectively blend data from both event cameras and frame-based cameras. The model operates in a two-stage architecture inspired by the Event deblurring physical model [46], using bottom-up attention to highlight pertinent features from the event stream for accurate deblurring. This bottom-up method is driven by the event cameras' data, which record intensity changes with high temporal resolution, enabling the model to focus on crucial motion information for deblurring. Additionally, the paper introduces the REBlur dataset, featuring real-world event streams paired with corresponding blurry and sharp images, to assess EFNet's performance in challenging scenarios. This study marks a significant advancement in motion deblurring by effectively utilizing event data, following bottom-up approach.

Gesture recognition is another application where attention using event data proves to be highly effective. Gruel's 2022 study [22] demonstrate the advantages of using neuromorphic attention for event data for recognizing hand and arm move-

ments. Gruel's method involves a spiking neural network that dynamically adapts to incoming event streams, focusing on regions of high activity that correspond to meaningful areas while filtering out irrelevant background noise. The regions with the highest density of event data are defined as the focal points of attention (RoI) within the respective model, enabling the system to prioritize areas with significant activity while discarding other unrelevant locations.

In the context of object recognition from a static camera, attention using event data offers certain advantages, especially in scenarios involving dynamic environments and varying lighting conditions. In Bulzomi's 2023 study [5], pedestrian detection is explored using a static camera setup equipped with an Event sensor. The dataset includes various recordings where event data is processed to filter out background noise and focus on the movement of pedestrians. This method leverages the asynchronous nature of event data to detect changes in the scene, allowing the system to efficiently isolate and recognize moving objects, such as pedestrians, against a static background. The definition of attention in Bulzomi's study aligns with Gruel's work [22], where RoI is determined as the position that contains the highest density of events, ensuring that the system concentrates on the most pertinent data.

Cancini's 2019 research [7] presents a novel approach for object recognition utilizing event cameras, focusing on enhancing processing efficiency in dynamic environments. The authors propose two models: the first model tracks event activity to identify RoI using a peak detection algorithm, implementing a top-down approach to focus computational efforts on the most active areas. The second model adapts the DRAW-based neural architecture [21] to handle event data effectively. These RoIs are then processed by a Phased LSTM recognition network, which combines spatial and temporal information for accurate object classification. The significant contribution of this research lies in the adaptation of attention mechanisms specifically for event cameras, demonstrating improved handling of translation and scale variations compared to conventional models.

Kong et al. (2024) [30] introduce a model for Event semantic segmentation, leveraging the pre-trained knowledge from image and text domains using CLIP [47]. The objective is to address the challenges of dense annotations and scalability in Event vision tasks by transferring semantically rich CLIP knowledge to event streams. The model employs a top-down approach, using text prompts to guide the attention to relevant regions of the event data, facilitating zero-shot segmentation. This top-down mechanism effectively aligns event features with textual descriptions, allowing for semantic coherence in the segmentation process.

Here again, most attention approaches for event data focus on bottom-up mechanisms, with few incorporating top-down attention guided by task relevance. Amongst the state-of-the-art papers described in this section and the previous section, we notice that the majority of work mainly deal with bottom-up attention, possibly because it is harder and not straightforward to model top-down attention.

3 Methodology

As outlined in subsection 2.2, this internship will focus on cognitive attention models, specifically incorporating both bottom-up and top-down attention mechanisms. Bottom-up attention is driven by data and focuses on automatic feature extraction from the data itself, while top-down attention involves predefined tasks and leverages higher-level guidance.

For bottom-up attention, we will use the YOLO (You Only Look Once) model [50], specifically YOLOv8, because the approach of YOLO allows this model to learn and adapt from the data without requiring predefined instructions. YOLO is a popular real-time object detection model known for its speed and accuracy. YOLOv8, the latest iteration by Ultralytics, improves upon previous versions with enhanced performance and efficiency. However, YOLOv8 was originally trained on frame-based image datasets (such as COCO), which may not be directly compatible with event images. To address this, YOLOv8 will be fine-tuned on event images to evaluate its performance with this unique type of data.

For top-down attention, we will utilize YOLO-World [8], an open-vocabulary model based on YOLO, because this model offers the flexibility to specify the classes to be detected before model training, making it a powerful tool for tasks requiring predefined object detection. This is particularly advantageous for autonomous driving scenarios, where the focus may shift depending on the environment—such as prioritizing "car" detection on highways or "pedestrian" detection in city centers. YOLO-World's top-down approach enables this task-specific focus. Additionally, YOLOWorld has the capability to predict new classes that were not part of the original training or fine-tuning dataset. This feature is crucial because it allows for greater flexibility and adaptability in real-world applications, where encountering unforeseen objects or scenarios is common. Like YOLO, YOLO-World is also initially trained on frame-based images, necessitating fine-tuning on event data to optimize its performance.

3.1 YOLO – Bottom-up attention model

YOLO (You Only Look Once) [50] represents a significant advancement in object detection by streamlining the detection process into a single-stage framework. Unlike older models that employ complex region proposal networks or multiple stages of processing, YOLO approaches object detection as a direct regression task. This design allows the model to examine the entire image in one go using a convolutional neural network (CNN), which enables fast and efficient detection in real-time scenarios. Figure 3 illustrates the architecture of the YOLO model.

In its architecture, YOLO splits the input image into a grid, with each grid cell responsible for predicting a set of bounding boxes, along with confidence scores and class probabilities. The model's architecture is influenced by the GoogLeNet design, featuring 24 convolutional layers followed by fully connected layers, which help in

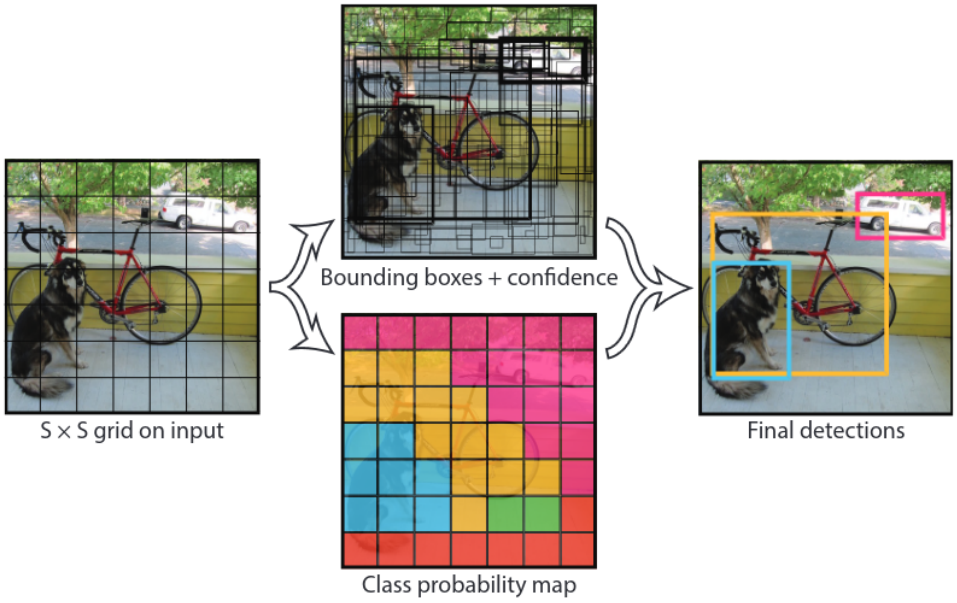

Fig. 3: YOLO Model. Source: [50]

accurately predicting the objects in the image. An important component of YOLO's detection strategy is the Intersection over Union (IoU) metric, which quantifies the overlap between predicted bounding boxes and the actual object locations. This metric is crucial for ensuring that the model can precisely locate objects within the image. YOLO's integrated approach allows it to consider the entire image context during both training and prediction, which reduces the likelihood of false positives from the background and enhances its ability to generalize across different datasets. The latest iteration, YOLOv8, builds on this foundation, offering improvements that make it particularly well-suited for modern applications in areas like autonomous vehicles, robotics, and real-time surveillance.

3.2 YOLO-World – Top-down attention model

YOLO-World [8] represents a significant advancement in object detection by addressing the limitations of traditional YOLO models, which are confined to a set of predefined categories. Building on the foundational YOLOv8 architecture, YOLO-World introduces open-vocabulary detection capabilities through the integration of vision-language modeling. The model employs a Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN) to merge visual features with text embeddings, allowing it to identify objects that go beyond the fixed categories typically supported by earlier YOLO versions. This capability is powered by a text encoder, pre-trained with CLIP, which converts input text into embeddings. These embeddings then interact with visual features in the network, enhancing the model's ability to associate

visual objects with their corresponding textual descriptions. Figure 4 provides an overview of the YOLO-World architecture.

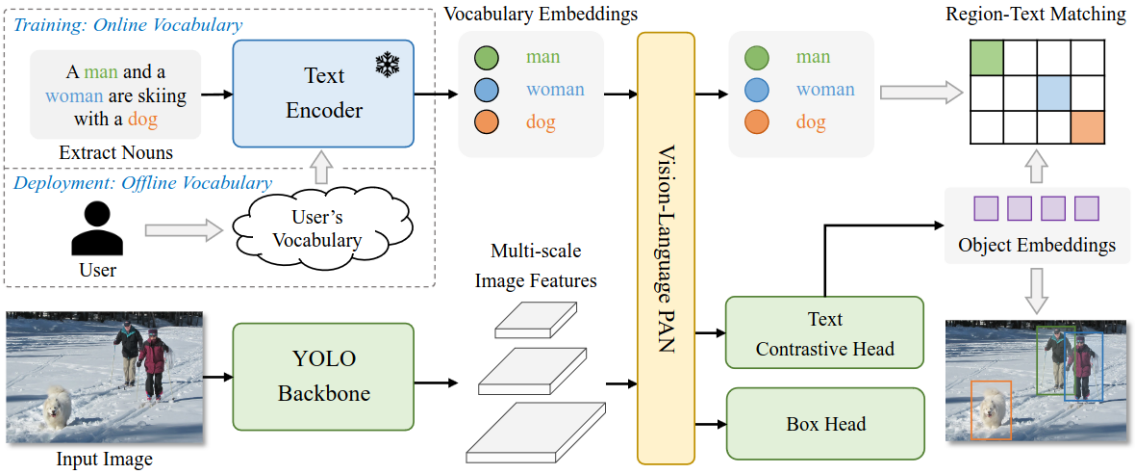

Fig. 4: YOLO-World Architecture. Source: [8]

The YOLO-World architecture includes a backbone for feature extraction, a path aggregation network for multi-scale feature refinement, and a detection head that handles bounding box regression and object classification. The model utilizes a region-text contrastive loss to align regions of the image with the provided textual input, a key feature that enables zero-shot detection. This means YOLO-World can detect objects that were not part of its training set, based solely on the textual cues provided during inference. Additionally, the model uses an Intersection over Union (IoU) loss to fine-tune the accuracy of bounding box predictions, ensuring precise localization of detected objects. With these innovations, YOLO-World delivers a robust solution for open-vocabulary object detection, making it well-suited for complex and dynamic real-world applications.

3.3 Finetune models

During the fine-tuning of YOLO and YOLO-World, various combinations of hyperparameters—such as epoch count, batch size, and image size—were tested to identify the most effective configuration. Since this internship primarily focuses on top-down attention using YOLO-World, a more detailed comparison of different hyperparameter combinations will be provided in the following section. For YOLO, we will present the best results obtained using the optimal hyperparameter settings in Section 5.

3.4 Evaluation metrics

The evaluation process leans on the confusion matrix, power consumption and other pivotal metrics commonly used in object detection, particularly for YOLO and YOLO-World:

Confusion matrix: The evaluation process begins with the confusion matrix, which is fundamental for understanding the performance of the YOLO and YOLO-World models. The confusion matrix is typically structured as a square matrix with dimensions corresponding to the number of classes in the classification task. For a multi-class classification problem, the matrix expands to include rows and columns for each class, showing the counts of correct and incorrect predictions for each class. The confusion matrix is a powerful diagnostic tool that helps to understand the performance of a classification model beyond mere accuracy, by highlighting specific areas where the model may be making errors, such as misclassifications of certain classes or a tendency towards false positives or negatives.

Precision: Precision measures the accuracy of positive predictions. It is the ratio of true positive detections to the total number of positive predictions. Precision indicates how many of the detected objects are relevant.

$$\textbf{Precision} = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Positives}} \tag{1}$$

Recall: Recall measures how well the model identifies all relevant objects. It is the ratio of true positive detections to the total number of actual positive objects. Recall indicates how many of the actual objects were detected.

$$\textbf{Recall} = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}} \tag{2}$$

mAP50 (Mean Average Precision at IoU=0.5): mAP50 is the average precision across all classes, calculated with a threshold Intersection over Union (IoU) of 0.5. It measures how well the model performs in detecting objects with a moderate overlap threshold.

mAP50-95 (Mean Average Precision at IoU thresholds from 0.5 to 0.95): mAP50-95 calculates the average precision across different IoU thresholds (from 0.5 to 0.95 in steps of 0.05). It provides a more comprehensive evaluation of the model's performance by considering various levels of overlap.

F1-Confidence Curve: This curve plots the F1 score (a harmonic mean of precision and recall) against different confidence thresholds. It helps in evaluating how precision and recall trade off as the confidence threshold for detections is varied.

$$\textbf{F1 Score} = 2 \cdot \frac{\textit{Precision} \cdot \textit{Recall}}{\textit{Precision} + \textit{Recall}} \tag{3}$$

Recall-Confidence Curve: This curve shows how recall changes with varying confidence thresholds. It helps in understanding how the recall rate (detection of

relevant objects) is affected as the model's confidence threshold for considering a detection is adjusted.

Precision-Recall Curve: This curve plots precision against recall for different confidence thresholds. It provides a graphical representation of the trade-offs between precision and recall for the model and is useful for evaluating the model's performance across different threshold settings.

Power consumption: Besides the metrics, another key consideration during this internship is the power consumption associated with using the model to predict test images. This aspect is particularly important for future real-world applications, where energy efficiency is crucial during the embedding process. The inference power consumption (or energy inference consumption) is calculated as following: the process begins by measuring the inference time per image in milliseconds. The GPU and CPU power usage in watts (W) are then recorded during inference. This recorded time is converted from milliseconds to hours. The total power usage (GPU + CPU) is multiplied by the converted time to calculate the energy consumption in kilowatt-hours (kWh) and then the energy can be converted to joules (J) by multiplying the kWh value by 3.6×10^6. If inference is performed in batches, the total energy is divided by the batch size to determine the per-image energy consumption. This method provides the energy consumption for each image.

The detailed evaluation of these results, including performance across all classes, will be thoroughly discussed in Section 5.

4 The data

4.1 Dataset description overview

The objective of this internship is to work with event data, specifically focusing on object detection for autonomous driving scenarios. Among the various event datasets discussed in Section 2, the DSEC-Detection dataset [1] is chosen due to its large size, complexity, and recent development. This dataset is known for its high quality and has been increasingly featured in recent research and publications, making it a valuable resource for studying and implementing attention models. Throughout the internship, attention models will be applied to the DSEC-Detection dataset to explore and advance object detection capabilities in autonomous driving contexts.

Original DSEC Dataset

The DSEC (Stereo Event Camera Dataset for driving scenarios) [19] [20], is developed by the Robotics and Perception Group led by Prof. Davide Scaramuzza at the University of Zurich. By its unique approach, DSEC is the first high-resolution, large-scale stereo dataset with event cameras. It is important to note that in this internship, we will not be addressing the stereo aspect of the dataset.

This dataset uniquely combines data from two high-resolution monochrome event cameras and two color frame cameras. The event cameras, with a resolution of

640×480 and a baseline of 60cm, do not have traditional exposure settings but adjust sensitivity according to lighting conditions, enabling detailed capture in both daylight and night conditions. Next to each event camera, there is a color FLIR Blackfly S USB3 camera with a resolution of 1440 × 1080 and a baseline of 51cm, operating in auto exposure mode to prioritize exposure time adjustments over increasing analog gain, thus reducing noise and minimizing motion blur at night. The dataset also includes LIDAR data and RTK GPS measurements, all synchronized through a microcontroller system. In DSEC, there are 53 sequences captured in different illumination conditions and offering ground truth disparity maps for developing and evaluating stereo vision algorithms.

DSEC-Detection dataset

DSEC is widely applied into various computer vision aims such as segmentation, detection or optical flow. In the scope of this internship, the work focuses on DSEC-Detection dataset, an extension specifically tailored for enhancing object detection and tracking capabilities in driving scenarios.

DSEC-Detection is quite large and complex dataset with size of 578Gb, comprising detection labels for 60 (53 sequences from the original DSEC dataset as well as 7 additional challenging scenarios), totaling 70,379 frames and 390,118 bounding boxes. For further studies, these 60 sequences are divided into 47 sequences for trainset and 13 sequences for testset. Each sequence in the dataset includes calibration data, events, images, and object detection labels. The setup comprises two event cameras, each paired with a neighboring color camera to capture information from both the "right" and "left" sides. However, after synthesizing and reconstructing the event data (as well as the frame images) from all cameras, the object detection process is primarily focused on data from the 'left' side. This consolidated approach ensures a streamlined analysis and simplifies the handling of visual information for object detection tasks. The detection labels are generated using a sophisticated object tracking algorithm on rectified image frames, which are then adjusted to fit the event camera frame perspective. Each detection is manually verified to remove inaccuracies and false tracks, ensuring high-quality and reliable data for analysis. The labels include bounding box coordinates, class information, and track identities, essential for detailed object-tracking studies. The detected classes comprise 8 categories: pedestrian, rider, car, bus, truck, bicycle, motorcycle, and train, making this dataset particularly valuable for research focused on dynamic and varied urban environments.

In summary, DSEC is a valuable dataset for exploring and developing stereo vision systems with event cameras, while DSEC-Detection dataset extends this framework to support workings in object detection and tracking. Together, these datasets offer invaluable tools for autonomous driving technologies and other applications where precise, real-time visual data is critical. Figure 5 [1] presents an example of DSEC-Detection visualization with gray-scale image, with event representation (in blue and red plots) and object label detection.

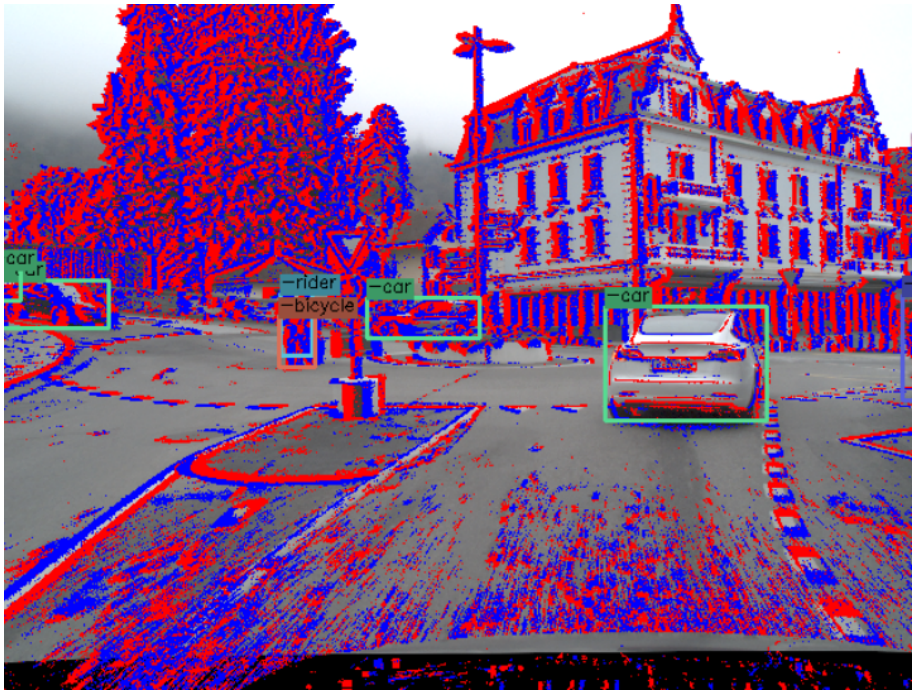

Fig. 5: DSEC-Detection visualization example. Source: [1]

4.2 Data preprocessing

The DSEC-Detection dataset provides event data in the format (x, y, t, p), where x and y are the coordinates, t is the timestamp, and p represents the polarity of detected light changes. For this internship, we use these event data to generate event images and corresponding labels, which are crucial for fine-tuning the attention models. The dataset includes 47 sequences for training and 13 sequences for testing. The data processing involves several key steps:

1. **Dataset Splitting:** We maintain the test set as is, while the training set is divided into 80% for training and 20% for validation. Each subset contains event images and labels for object detection.
2. **Label Cleaning:** YOLOv8 and YOLO-World require labels with 5 values (class, x_{center}, y_{center}, width, height). DSEC-Detection labels include 6 values (class, x_{center}, y_{center}, width, height, probability). We remove the probability values to match the required format.
3. **Prompt Generation:** YOLO-World requires text prompts describing each image, which are not provided in DSEC-Detection. We automatically generate prompts in the format "This image contains ..." with details on the number of each class present.
4. **Missing Value Removal:** Any missing values in the training, validation, and test datasets are removed to ensure data integrity.

5. **Out-of-Bounds Values:** Labels must adhere to YOLOv8 and YOLO-World specifications: one row per object, formatted as class x_{center} y_{center} width height, with coordinates normalized between 0 and 1. Labels with out-of-bounds values are corrected or removed.

The figure 6 illustrates the class distribution across the three datasets (training, validation, and test) following data preprocessing. As depicted, the "car" class is the most prevalent, followed by "pedestrian" as the second most common class. In contrast, classes such as "bicycle" and "motorbike" have significantly fewer instances. This distribution aligns with real-world expectations, as the data was collected along a road frequented by cars, resulting in a higher number of car instances compared to other classes.

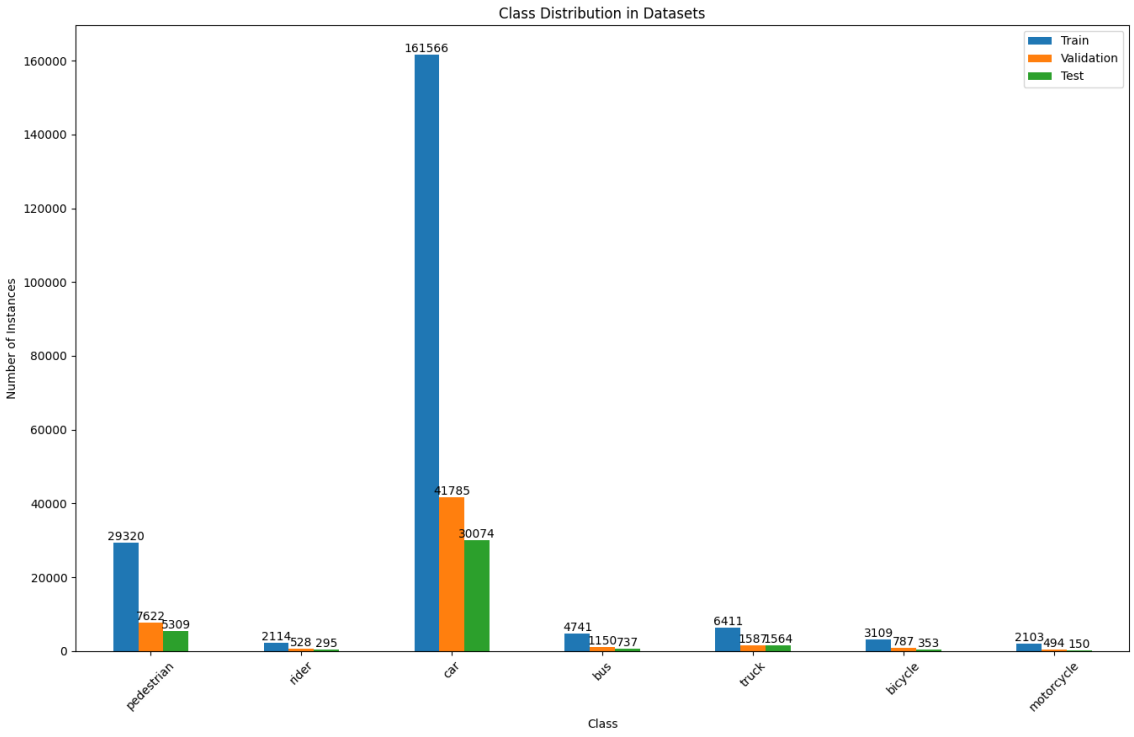

Fig. 6: Class distribution

5 Results and discussion

In this section, we will present the detection results obtained by fine-tuning YOLO and YOLO-World on the DSEC-Detection dataset. We will discuss the selection of the optimal hyperparameter combinations and conclude with a comparison of the results from both models, providing explanations for the observed differences in

performance. It is important to emphasize that, due to the novelty of applying top-down attention models to event data, this internship primarily focused on adapting the YOLO-World model to this data type. As for YOLO, we will briefly touch on some key results and metrics. Additionally, it is crucial to note that the decision to plot the event images using blue and red points does not impact the results.

As mentioned in Section 4, the "car" class is the most prevalent and serves as the primary focus for model training during this internship. Therefore, the model's performance will be primarily assessed using the confusion matrix and other metrics, with a particular emphasis on the "car" class.

5.1 YOLOv8

5.1.1 Results To provide more insight into the detection outcomes, Figure 7 displays an example with the labeled image on the left and YOLO's detection results on the right. After fine-tuning on event images, YOLO, originally trained on frame-based images, demonstrated a marked improvement in its ability to recognize objects in the new data format. Notably, YOLO detected objects in event data that are challenging for the human eye to identify, given our familiarity with frame-based images. This is a particularly noteworthy outcome of the experiment.

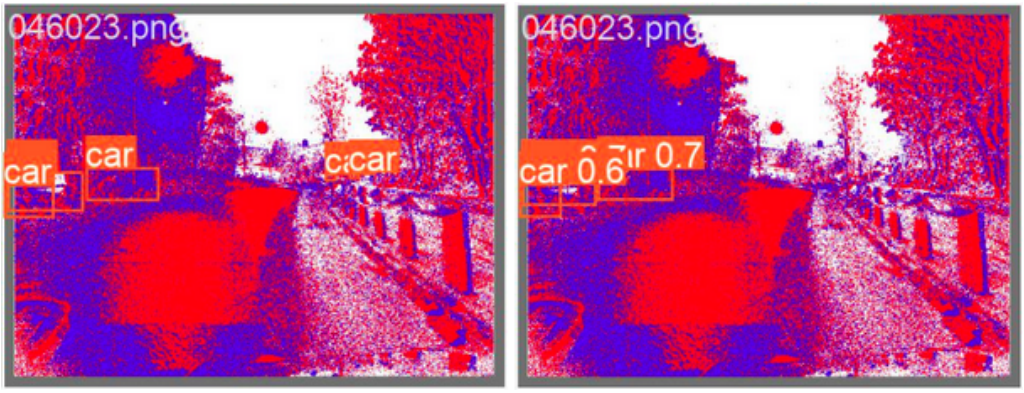

Fig. 7: Detection example - YOLO for DSEC-Detection

The optimal performance of YOLO on the DSEC-Detection dataset was achieved with 100 epochs, a batch size of 32, and an image size of 640. As shown in the normalized confusion matrix (Figure 8), the model successfully detected all classes with relatively high accuracy, particularly for the "car" class, which achieved an accuracy of 0.75. Even the less frequent classes, such as truck, bicycle, and motorbike, demonstrated commendable accuracy levels of 0.76, 0.76, and 0.72, respectively.

Despite these high accuracy figures, other key metrics like Precision and Recall were not as strong. For all classes, the Precision was 0.154 and the Recall was 0.584, while for the "car" class specifically, these values were 0.172 and 0.661, respectively.

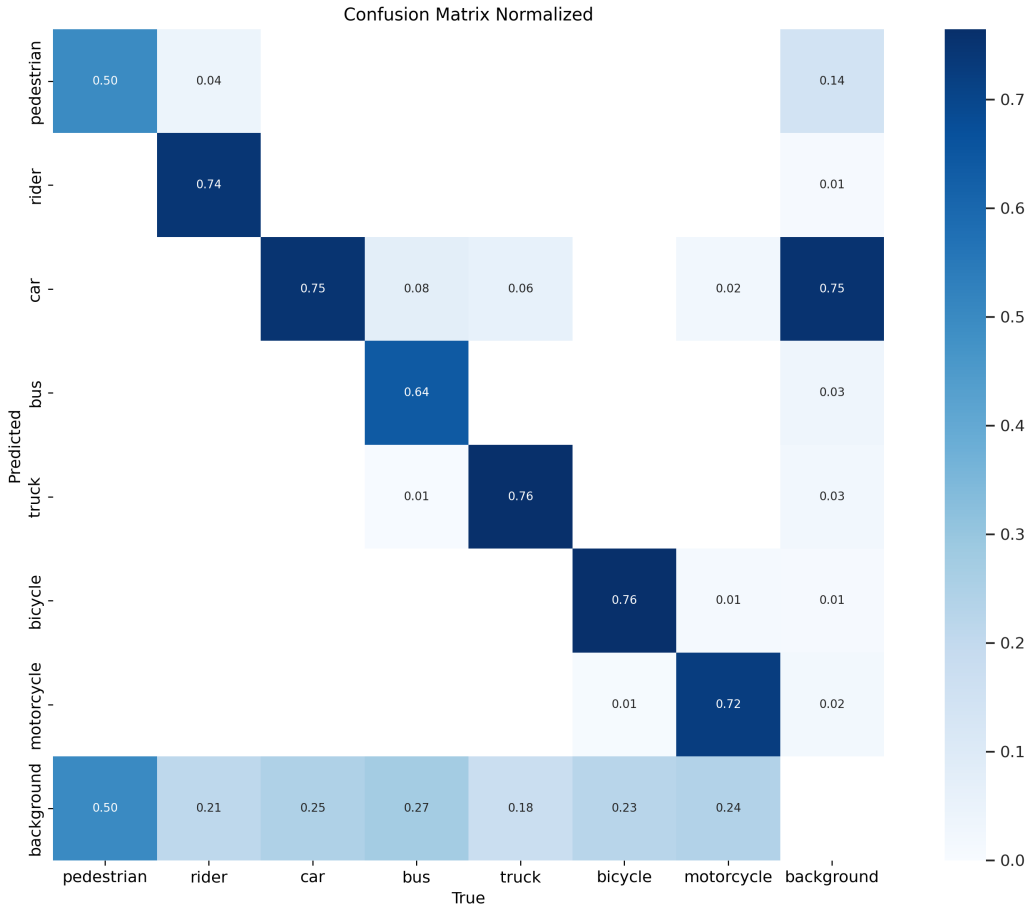

Fig. 8: Confusion matrix normalized - YOLO for DSEC-Detection

The mAP50 and mAP50-95 scores were 0.155 and 0.107 for all classes, and 0.156 and 0.11 for the "car" class, with similar patterns observed across other classes. The details for metrics of YOLO is presented in Table 2. These results will be explored further in the discussion section 5.1.2.

5.1.2 Discussion Several factors could explain the observed high accuracy but lower Precision, Recall, and other metrics. One major factor is the background class bias: in YOLO, any part of the image not identified as an object is categorized as "background." In event images, where objects are represented by red and blue points (indicating positive and negative events), it becomes challenging to distinguish between the background and other object classes. This overwhelming representation of the background could bias the model towards predicting this class more frequently, resulting in high accuracy but lower performance metrics (Precision, Recall, mAP50, mAP50-95) for the actual object classes.

Table 2: YOLO Metrics Summary

Class	Precision	Recall	mAP50	mAP50-95
all	0.154	0.584	0.155	0.107
pedestrian	0.141	0.321	0.106	0.0608
rider	0.142	0.667	0.163	0.11
car	0.172	0.611	0.156	0.11
bus	0.15	0.565	0.14	0.105
truck	0.176	0.681	0.165	0.122
bicycle	0.145	0.639	0.151	0.0971
motorcycle	0.155	0.602	0.202	0.143

Moreover, the relatively low number of instances for other classes like "truck," "bicycle," and "motorbike" further limits the model's ability to learn and accurately predict these categories. The insufficient representation of these classes in the dataset likely contributes to the reduced effectiveness of the model in correctly identifying them. Addressing this imbalance could be a promising direction for future research, potentially involving techniques like data augmentation or the inclusion of additional labeled data to improve the model's performance across all object classes.

5.2 YOLO-World

5.2.1 Results The primary focus of this internship is the application of top-down attention within the relatively unexplored domain of event data. Leveraging YOLO-World, a novel approach in this context, offers promising results and sets the stage for future research endeavors. As discussed at the beginning of Section 3, YOLO-World is a recent object detection model that allows users to pre-define the classes they wish to focus on, embodying the principles of top-down attention. This capability to "control" the detection output by selecting specific classes of interest makes it particularly appealing for autonomous driving scenarios, where a driver's attention may need to shift depending on various factors such as road conditions, weather, or crowded environments.

YOLO-World not only allows for the selection of specific classes of interest for prediction, but it can also identify new classes that were not included in the fine-tuning dataset. For instance, when fine-tuning YOLO-World with event data, only seven classes were used: "pedestrian," "rider," "car," "bus," "truck," "bicycle," and "motorcycle." In the example provided, the same moment is captured in an RGB image (Figure 9), the corresponding event image (Figure 10), and the predictions made by the fine-tuned YOLO-World model (Figure 11). Remarkably, in Figure 11, the model predicts a new class, "person," even though this class was not part of the original seven classes. YOLO-World's ability to predict a new class is due to its advanced vision-language model that leverages the relationship between visual features and textual descriptions, allowing it to generalize beyond the specific classes it was trained on. In the following parts, further discussion will focus on the metrics and

power consumption of YOLO-World in object detection using the DSEC-Detection dataset.

Fig. 9: Frame-based image

Fig. 10: Event image

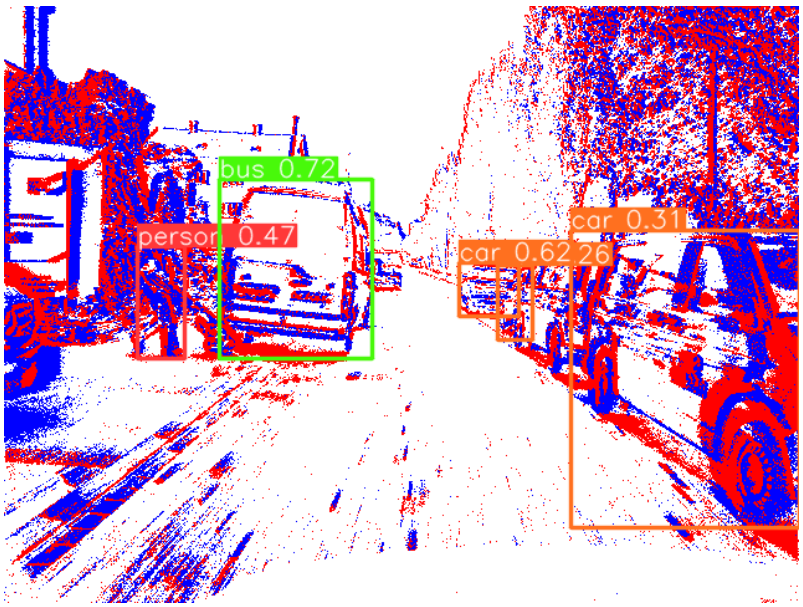

Fig. 11: Prediction of YOLO-World

Table 3 presents the comparison of combinations of hyperparameters for fine-tuning model YOLO-World with the DSEC-Detection dataset. Given the constraints of computational resources and the time limitations of this internship, the optimal results were achieved with an image size of 512, 20 epochs, and a batch size of 32. The training duration was 27.306 hours, with a total runtime (including training, validation, and testing) of 27.521 hours. The precision, recall, mAP50, and mAP50-95 scores were 0.594, 0.342, 0.374, and 0.248, respectively—higher than those achieved with YOLO. These metrics were even more pronounced for the dominant "car"

Table 3: YOLO-World Metrics – Comparison with different combination of hyperparameters

Image size	epoch	batch	P	R	mAP50	mAP50-95	P for car	R for car	mAP50-95 for car
256	7	32	0.442	0.214	0.220	0.132	0.648	0.527	0.347
512	7	32	0.596	0.343	0.371	0.239	0.714	0.673	0.492
512	20	32	0.594	0.342	0.374	0.248	0.718	0.657	0.489

class, with scores of 0.718, 0.657, 0.693, and 0.489, reflecting the model's stronger performance when sufficient data is available for training.

In addition to evaluating the standard metrics, we also considered the power consumption of each model, as shown in Table 4. For each model, the training time is measured in hours and represents the time required to fine-tune YOLOWorld using the training dataset. The total time, also in hours, includes the time spent on fine-tuning, validation, and testing. The inference time (measured in milliseconds) and power consumption (measured in joules) represent the average time and energy required to predict an image in the test dataset using the fine-tuned model. These metrics, particularly inference time and power consumption, are crucial for future real-world applications where energy efficiency is essential. The formula and process for calculating power consumption are detailed in Section 3.

Table 4: YOLO-World – Power consumption – Comparison with different combination of hyperparameters

Image size	epoch	batch	Time train (hour)	Time total (hour)	Reference Time (ms)	Power Consumption (J)
256	7	32	3.515	3.866	14.3	1.1154
512	7	32	10.760	11.273	98.3	7.469
512	20	32	27.306	27.521	27.0	2.025

Tables 3 and 4 reveal that the model fine-tuned with an image size of 512, 20 epochs, and a batch size of 32 delivers the best performance across metrics. However, this configuration also requires the longest training time, which is expected since increasing the image size and the number of epochs naturally extends the training duration. Notably, despite its longer training time, this model demonstrates a significantly lower inference time and reduced power consumption when predicting an image—only about one-third of the model trained with 512 image size, 7 epochs, and a batch size of 32. This reduction in inference time and power consumption is likely due to the model's improved efficiency and optimization after more extensive training. Additionally, while the model fine-tuned with 512 image size, 20 epochs, and a batch size of 32 shows slightly higher inference time and power consumption than the one trained with 256 image size, 7 epochs, and a batch size of 32, the

difference is not substantial. In conclusion, the optimal combination of hyperparameters, considering both performance metrics and power consumption, is the model fine-tuned with an image size of 512, 20 epochs, and a batch size of 32. The detailed of metrics for YOLO-World with best combination of hyperparameters is presented in Table 5.

Table 5: YOLO-World Metrics Summary with best combination of hyperparameters

Class	Precision	Recall	mAP50	mAP50-95
all	0.594	0.342	0.374	0.248
pedestrian	0.615	0.41	0.441	0.264
rider	0.761	0.273	0.394	0.245
car	0.718	0.657	0.693	0.489
bus	0.719	0.218	0.261	0.208
truck	0.597	0.371	0.401	0.252
bicycle	0.615	0.363	0.383	0.257
motorcycle	0.136	0.103	0.0445	0.022

Additionally, Figure 12 presents the normalized confusion matrix for YOLO-World with the best combination of hyperparameters. We can see that the accuracy for all classes detected by YOLO-World is reduced in comparison to the detection by YOLO. This could be explained in the following part 5.2.2.

5.2.2 Discussion The precision, recall, mAP50, and mAP50-95 scores, while not exceedingly high, surpass those obtained with YOLO. This improvement can be attributed to YOLO-World's use of image descriptions, which provide additional context and guidance during object detection, enhancing the model's performance. The higher metrics for the "car" class are particularly noteworthy, as this class has the most instances available for training, allowing the model to learn and predict more effectively.

As observed in Table 3, increasing the image size and the number of epochs generally improves model performance. However, due to the constraints of time and computational resources, the best achievable model in this internship used an image size of 512 with 20 epochs. For future research, we anticipate that extending the training period and employing larger image sizes could yield even better results, allowing for more comprehensive fine-tuning and improved detection accuracy.

The difference in performance metrics between YOLO and YOLO-World on the same dataset, where YOLO has higher accuracy in the confusion matrix but YOLO-World has higher metrics like precision, recall, mAP50, and mAP50-95, can be explained by several factors related to how each model is designed, how they handle predictions, and what their primary strengths are:

1. **Model Architecture:** YOLO is optimized for speed and may sacrifice some accuracy, especially in object localization. YOLO-World, with its multimodal

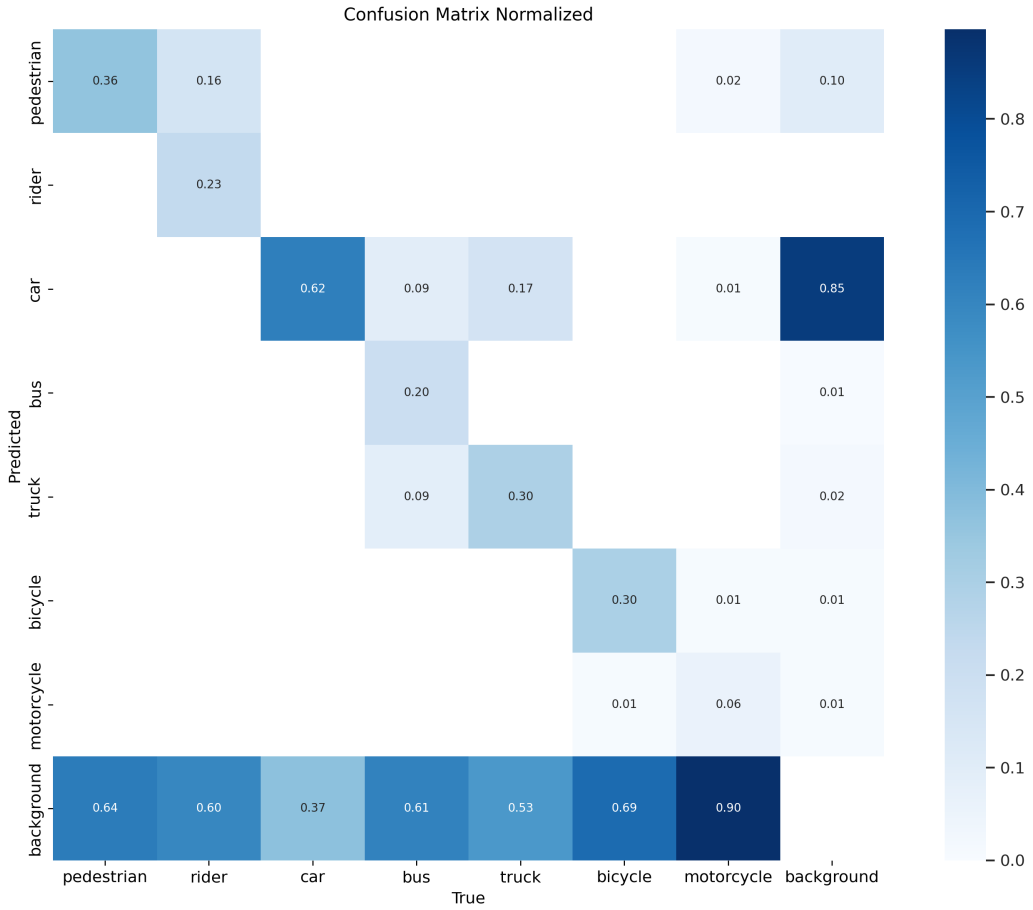

Fig. 12: Confusion matrix normalized - YOLO-World for DSEC-Detection

approach, uses text prompts to improve accuracy in distinguishing between objects and background, leading to better precision, recall, and mAP.

2. **Metric Focus:** YOLO might show higher accuracy because it correctly predicts dominant classes more often, but this doesn't always translate to better precision or recall. YOLO-World, on the other hand, is optimized to make more accurate and higher-quality predictions across all classes, reflected in higher precision, recall, and mAP values.
3. **Generalization:** YOLO may overfit to training data, leading to higher accuracy but lower generalization. YOLO-World's use of additional context (like text prompts) helps it generalize better, improving overall metrics despite potentially lower raw accuracy.

In summary, YOLO might achieve higher accuracy because it is better at making correct predictions for the dominant class, but it might struggle with precision and recall across all classes. YOLO-World leverages additional modalities (like text

prompts) and more sophisticated model architecture, leading to better overall precision, recall, and mAP, which are more indicative of high-quality object detection performance across all classes.

6 Conclusion

The exploration of attention mechanisms tailored to event data, particularly within the context of autonomous driving, represents a significant step forward in the development of efficient and responsive computer vision systems. Traditional computer vision models have predominantly relied on frame-based images, which, while effective, often generate excessive amounts of redundant data and demand substantial computational resources. Event cameras, on the other hand, offer a promising alternative by capturing only the changes in brightness at each pixel, thereby minimizing data redundancy and reducing power consumption. This ability to focus on dynamic changes in the environment makes event cameras particularly suitable for real-time applications where responsiveness and efficiency are crucial.

The objective of the internship project has focused on understanding and enhancing the role of attention mechanisms—both bottom-up and top-down—in processing event data. While bottom-up attention models have been more widely studied and implemented in existing computer vision systems, top-down attention models remain relatively unexplored, especially in the context of event data. This research has made significant progress in bridging this gap, both in terms of theoretical exploration and practical implementation.

The key contributions of the internship include a comprehensive survey on event data and attention models, as well as the practical application of these insights to detection tasks using the DSEC-Detection dataset. This survey, which is currently accepted for a poster presentation in ICONIP 2024 Conference [3], offers a detailed analysis of the state-of-the-art in attention models and event data, serving as a crucial resource for researchers looking to advance the capabilities of computer vision systems. One of the primary objectives of this internship was to adapt existing models, specifically YOLO and YOLO-World, to work with event data. YOLO is chosen as the bottom-up attention model for its data-driven approach and its speed and efficiency in real-time object detection, while YOLOWorld was selected as the top-down attention model due to its ability to incorporate predefined tasks and detect new classes using text prompts. Both models were originally designed for frame-based images, necessitating fine-tuning and adaptation to work effectively with event data.

Results YOLO and YOLOWorld have been demonstrated as bottom-up and top-down attention model respectively. The results of this adaptation process were promising, particularly in demonstrating the potential of attention models in handling event data. YOLO-World, with its ability to leverage text prompts and inte-

[3] ICONIP 2024, URL: https://iconip2024.org/

grate contextual information, showed improvements in precision, recall, and mAP (Mean Average Precision) compared to the traditional YOLO model. This is particularly noteworthy given the challenges associated with event data, such as its sparse and asynchronous nature, which often complicates the application of traditional computer vision models. The higher metrics achieved by YOLO-World suggest that top-down attention, when properly implemented, can greatly enhance the accuracy and efficiency of event-based vision systems. Additionally, the detection results reveal its capability to identify new classes not included in the original seven classes of the event dataset. Furthermore, the model's power consumption is also taken into consideration, highlighting its suitability for practical applications. The findings from this internship have important implications for the future of computer vision research, particularly in the context of autonomous driving and other real-time applications. The integration of top-down and bottom-up attention mechanisms has the potential to greatly enhance the efficiency and accuracy of vision systems, allowing them to operate more effectively in dynamic and complex environments. This is especially relevant in scenarios where quick and accurate decision-making is critical, such as in autonomous vehicles navigating through unpredictable road conditions.

Challenges and future research proposals The research conducted during this internship also highlighted several challenges and areas for improvement. For instance, the performance of the models was heavily influenced by the class distribution within the dataset, with more frequent classes like "car" achieving higher accuracy and mAP scores compared to less frequent classes such as "bicycle" or "motorcycle." This imbalance in the dataset limited the models' ability to learn and accurately predict less common classes, suggesting a need for more balanced datasets or advanced data augmentation techniques to enhance the robustness of the models. Moreover, the distinction between background and object classes in event data proved to be a significant challenge. Event images, characterized by their representation of objects as clusters of red and blue points corresponding to positive and negative events, often led to difficulties in distinguishing between background and foreground objects. This issue likely contributed to the lower performance metrics observed in some cases, despite the overall high accuracy of the models. Addressing this challenge will be crucial for further improving the performance of event-based vision systems. Instead of automatically generated prompts, manually crafted, precise, and detailed descriptions should be employed to enhance accuracy. Additionally, developing more refined and comprehensive datasets will further boost model effectiveness. Addressing data imbalances through advanced preprocessing techniques will also be crucial. These enhancements are expected to significantly advance the performance and reliability of attention models for event data, particularly in applications like autonomous driving.

In conclusion, this internship has made contributions to the understanding and application of attention mechanisms in event-based vision systems. The adaptation

of YOLO and YOLO-World models to event data has demonstrated the potential of top-down attention mechanisms in enhancing the performance of these systems, particularly in dynamic and real-time environments like autonomous driving. While there are still challenges to be addressed, the findings from this research provide a strong foundation for future advancements in the field. Continued exploration and refinement of attention mechanisms, combined with advances in neuromorphic computing, hold the promise of pushing the boundaries of what is possible in computer vision, paving the way for more efficient, accurate, and adaptable vision systems that can meet the demands of increasingly complex real-world applications.

Acknowledgments

This work was supported by the project NAMED (ANR-23-CE45-0025-01) of the French National Research Agency (ANR). This work was supported by the French government through the France 2030 investment plan managed by the National Research Agency (ANR), as part of the Initiative of Excellence Université Côte d'Azur under reference number ANR- 15-IDEX-01.

The sub-section 2.2.1 is supported by Laurent Sparrow [4], Université de Lille, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives.

[4] Laurent SPARROW, laurent.sparrow@univ-lille.fr

References

1. DSEC-Detection. Online, https://dsec.ifi.uzh.ch/dsec-detection/
2. Prophesee's Event-Based Camera Reaches High Resolution. Online, https://spectrum.ieee.org/prophesees-eventbased-camera-reaches-high-resolution
3. Almatrafi, M., Baldwin, R., Aizawa, K., Hirakawa, K.: Distance surface for event-based optical flow. IEEE transactions on pattern analysis and machine intelligence **42**(7), 1547–1556 (2020)
4. Binas, J., Neil, D., Liu, S.C., Delbruck, T.: Ddd17: End-to-end davis driving dataset. arXiv preprint arXiv:1711.01458 (2017)
5. Bulzomi, H., Gruel, A., Martinet, J., Fujita, T., Nakano, Y., Bendahan, R.: Object detection for embedded systems using tiny spiking neural networks: Filtering noise through visual attention. In: 2023 18th International Conference on Machine Vision and Applications (MVA). pp. 1–5. IEEE (2023)
6. Calabrese, E., Taverni, G., Awai Easthope, C., Skriabine, S., Corradi, F., Longinotti, L., Eng, K., Delbruck, T.: Dhp19: Dynamic vision sensor 3d human pose dataset. In: CVPR workshops. pp. 0–0 (2019)
7. Cannici, M., Ciccone, M., Romanoni, A., Matteucci, M.: Attention mechanisms for object recognition with event-based cameras. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2019)
8. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16911 (2024)
9. Cheng, W., Luo, H., Yang, W., Yu, L., Chen, S., Li, W.: Det: A high-resolution dvs dataset for lane extraction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (2019)
10. Connor, C.E., Egeth, H.E., Yantis, S.: Visual attention: bottom-up versus top-down. Current biology **14**(19) (2004)
11. De Tournemire, P., Nitti, D., Perot, E., Migliore, D., Sironi, A.: A large scale event-based detection dataset for automotive. arXiv preprint arXiv:2001.08499 (2020)
12. Delbruck, T., et al.: Frame-free dynamic digital vision. In: Proceedings of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society. vol. 1, pp. 21–26. Citeseer (2008)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Engelke, U., Le Callet, P.: Perceived interest and overt visual attention in natural images. Signal Processing: Image Communication **39** (2015)
15. de Freitas, N.: Learning where to attend with deep architectures for image tracking. Neural Computation **24**(8) (2012)
16. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM Transactions on Applied Perception (TAP) **7**(1), 1–39 (2010)
17. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5633–5643 (2019)
18. Gehrig, D., Scaramuzza, D.: Pushing the limits of asynchronous graph-based object detection with event cameras. arXiv preprint arXiv:2211.12324 (2022)
19. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3) (2021)
20. Gehrig, M., Millhäusler, M., Gehrig, D., Scaramuzza, D.: E-raft: Dense optical flow from event cameras. In: 2021 International Conference on 3D Vision (3DV). IEEE (2021)
21. Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: Draw: A recurrent neural network for image generation. In: ICML (2015)
22. Gruel, A., Vitale, A., Martinet, J., Magno, M.: Neuromorphic event-based spatio-temporal attention using adaptive mechanisms. In: 2022 IEEE 4th international conference on artificial intelligence circuits and systems (AICAS). IEEE (2022)
23. Guo, S., Delbruck, T.: Low cost and latency event camera background activity denoising. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

24. Hu, Y., Binas, J., Neil, D., Liu, S.C., Delbruck, T.: Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–6. IEEE (2020)
25. Hu, Y., Liu, H., Pfeiffer, M., Delbruck, T.: Dvs benchmark datasets for object tracking, action recognition, and object recognition. Frontiers in neuroscience (2016)
26. Iacono, M., D'Angelo, G., Glover, A., Tikhanoff, V., Niebur, E., Bartolozzi, C.: Proto-object based saliency for event-driven cameras. In: IROS. IEEE (2019)
27. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews neuroscience **2**(3) (2001)
28. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI **20**(11), 1254–1259 (1998)
29. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023), https://github.com/ultralytics/ultralytics
30. Kong, L., Liu, Y., Ng, L.X., Cottereau, B.R., Ooi, W.T.: Openess: Event-based semantic scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
31. Le Callet, P., Niebur, E.: Visual attention and applications in multimedia technologies. Proceedings of the IEEE **101**(9), 2058–2067 (2013)
32. Lee, H., Kwon, H., Robinson, R.M., Nothwang, W.D., Marathe, A.M.: Dynamic belief fusion for object detection. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2016)
33. Li, H., Liu, H., Ji, X., Li, G., Shi, L.: Cifar10-dvs: an event-stream dataset for object classification. Frontiers in neuroscience **11**, 244131 (2017)
34. Li, J., Dong, S., Yu, Z., Tian, Y., Huang, T.: Event-based vision enhanced: A joint detection framework in autonomous driving. In: 2019 ieee international conference on multimedia and expo (icme). IEEE (2019)
35. Liang, Z., Chen, G., Li, Z., Liu, P., Knoll, A.: Event-based object detection with lightweight spatial attention mechanism. In: 2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM). IEEE (2021)
36. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128x128 120db 15mus latency asynchronous temporal contrast vision sensor. IEEE J. of solid-state circuits (2008)
37. Lindsay, G.W.: Attention in psychology, neuroscience, and machine learning. Frontiers in computational neuroscience **14** (2020)
38. Liu, M., Delbruck, T.: 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)
39. Lungu, I.A., Corradi, F., Delbrück, T.: Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In: 2017 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE (2017)
40. Maqueda, A.I., Loquercio, A., Gallego, G., García, N., Scaramuzza, D.: Event-based vision meets deep learning on steering prediction for self-driving cars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5419–5427 (2018)
41. Martinet, J., Lablack, A., Lew, S., Djeraba, C.: Gaze based quality assessment of visual media understanding. In: 1st International Workshop on Computer Vision and Its Application to Image Media Processing (WCVIM) in conjunction with the 3rd Pacific-Rim Symposium on Image and Video Technology (PSIVT), Tokyo-Japan (2009)
42. Miao, S., Chen, G., Ning, X., Zi, Y., Ren, K., Bing, Z., Knoll, A.: Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. Frontiers in neurorobotics **13** (2019)
43. Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. In: IROS. pp. 1–9. IEEE (2018)
44. Noudoost, B., Chang, M.H., Steinmetz, N.A., Moore, T.: Top-down control of visual attention. Current opinion in neurobiology **20**(2) (2010)
45. Orchard, G., Jayawant, A., Cohen, G.K., Thakor, N.: Converting static image datasets to spiking neuromorphic datasets using saccades. Frontiers in neuroscience **9** (2015)
46. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
48. Rai, Y., Le Callet, P., Cheung, G.: Quantifying the relation between perceived interest and visual salience during free viewing using trellis based optimization. In: 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). IEEE (2016)
49. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE transactions on pattern analysis and machine intelligence **43**(6), 1964–1980 (2019)
50. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
51. Scaramuzza, D.: Tutorial on event-based cameras. In: IROS 2015: Proc. of the 2nd Workshop on Alternative Sensing for Robot Perception (2015)
52. Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., Benosman, R.: Hats: Histograms of averaged time surfaces for robust event-based object classification. In: CVPR (2018)
53. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V.: Event-based fusion for motion deblurring with cross-modal attention. In: European conference on computer vision. Springer (2022)
54. Tsotsos, J.K.: A computational perspective on visual attention. MIT Press (2021)
55. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2022)
56. Vaishnav, M.: Exploring the role of (self-) attention in cognitive and computer vision architecture. Ph.D. thesis, Université Paul Sabatier-Toulouse III (2023)
57. Vasudevan, A., Negri, P., Di Ielsi, C., Linares-Barranco, B., Serrano-Gotarredona, T.: Sl-animals-dvs: event-driven sign language animals dataset. Pattern Analysis and Applications pp. 1–16 (2022)
58. Vasudevan, A., Negri, P., Linares-Barranco, B., Serrano-Gotarredona, T.: Introduction and analysis of an event-based sign language dataset. In: International Conference on Automatic Face and Gesture Recognition (2020)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
60. Wang, J., Chandler, D.M., Le Callet, P.: Quantifying the relationship between visual salience and visual importance. In: Human vision and electronic imaging XV. vol. 7527. SPIE (2010)
61. Weikersdorfer, D., Adrian, D.B., Cremers, D., Conradt, J.: Event-based 3d slam with a depth-augmented dynamic vision sensor. In: 2014 IEEE international conference on robotics and automation (ICRA). pp. 359–364. IEEE (2014)
62. Yang, Z., Mondal, S., Ahn, S., Xue, R., et al.: Unifying top-down and bottom-up scanpath prediction using transformers. In: CVPR (2024)
63. Yao, M., Gao, H., Zhao, G., Wang, D., Lin, Y., Yang, Z., Li, G.: Temporal-wise attention spiking neural networks for event streams classification. In: CVPR (2021)
64. Zheng, Y., Zemel, R.S., Zhang, Y.J., Larochelle, H.: A neural autoregressive approach to attention-based recognition. IJCV **113** (2015)
65. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)
66. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Ev-flownet: Self-supervised optical flow estimation for event-based cameras. arXiv preprint arXiv:1802.06898 (2018)
67. Zihao Zhu, A., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based optical flow using motion compensation. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)