Cognitive attention models for event data: survey and application

Anonymous WACV Algorithms Track submission

Paper ID *****

Abstract

Event cameras capture pixel-level changes in brightness asynchronously, offering high temporal resolution, low latency, and efficient data processing. Cognitive attention, a concept rooted in neuroscience and mimicking the human visual system, differs from the attention mechanism used in Transformer models in computer vision. Cognitive attention focuses on how the brain prioritizes and processes sensory input, whereas Transformer-based attention emphasizes feature weighting within neural networks. Although cognitive attention holds significant potential in computer vision, there are only a few studies on its implementation and application, with even fewer for event data, particularly top-down attention models, which remain novel. Furthermore, there is a lack of structured studies that fully connects the terminology and understanding of cognitive attention from neuroscience with its application in computer vision techniques. This paper addresses these gaps by first offering a comprehensive survey of existing cognitive attention models, following by a review of existing works for both RGB and event data. This survey serves as a foundation for future research. Additionally, we apply YOLO as a bottom-up model and YOLO-World as a top-down model to event data for object detection in autonomous driving scenarios, marking a promising step forward in this emerging field.

1. Introduction

In 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. Event cameras are a significant advancement over traditional frame-based cameras, capturing changes in brightness (*events*) at each pixel asynchronously [27]. This results in high temporal resolution, low latency, and reduced power consumption, making event cameras ideal for dynamic environments and high dynamic range conditions. Captured by event cameras, event data is represented as (x, y, t, p) where x and y are coordinates, t is the event timestamp, and p indicates brightness polarity. Event cameras and their unique data formats offer substantial potential for future research.

Besides event cameras and evant data, 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. Even though the application of these attention mechanisms are vast, there are only few existing studies of these into event data.

Our main contributions are as follows:

- We address the gap in understanding and terminology in cognitive attention for event data by providing a comprehensive survey which can be considered as a valuable resource for future research.
- We review existing studies using cognitive attention models for both RGB and event data for various computer vision applications, and classify them into either *top-down* or *bottom-up* attention.
- We apply YOLO [14] and YOLO-World [5] to the DSEC-Detection dataset [1], testing bottom-up and top-down attention models on a large event dataset for autonomous driving, resulting in very promising findings.

This study is part of an interdisciplinary collaborative research project, (HIDDEN FOR BLIND REVIEW)[1] , that aims to develop computational models of cognitive attention that combine both event sensors and RGB sensors, with improved power efficiency in embedded electronics, mimicking the human visual system to selectively focus on regions of interest.

The paper is structured as follows: Section 1 brings a general introduction for event data and attention models. Section 2 explores visual attention survey, focusing on

[1]Project number (HIDDEN FOR BLIND REVIEW)

bottom-up and top-down mechanisms for RGB and event data. The methodology, including dataset description, is detailed in Section 3. Section 4 presents and analyzes the results, while Section 5 concludes with recommendations for future research.

2. Attention for event data in computer vision

In our study, the term *attention* refers to *cognitive attention* in neuroscience, distinct from the Transformer architecture's attention mechanism [31]. This section presents a comprehensive survey about cognitive attention, covering both bottom-up and top-down approaches on both RGB and event data, offering a structured overview and guidance for future research in this domain.

2.1. Cognitive attention

Cognitive attention involves higher-order processes beyond sensory perception, such as decision-making and problem-solving [30]. These mechanisms enable the brain to prioritize and efficiently process sensory information from the environment. Tsotsos and colleagues define attention as "the process by which the brain controls and tunes information processing" [29]. Visual quality rapidly decreases from the center of gaze into a low-resolution surround, prompting constant eye movements (saccades) to redirect the fovea for stable visual information acquisition. Understanding these movements is crucial for understanding human visual processing. L. Itti and C. Koch (2001) [12] suggested that attention is selectively directed using both bottom-up, image-based cues, and top-down, task-dependent cues, as described in subsection 2.2.

2.2. Bottom-up and top-down attentions

Bottom-up attention, or exogenous attention, is driven by sensory input and operates rapidly and involuntarily [16] [7]. Saliency is a key concept, referring to the degree to which a stimulus stands out from its surroundings [25]. Saliency-driven attention is guided by the conspicuousness of features such as color, orientation, and motion, with salient stimuli attracting attention spontaneously. Computational models, such as saliency maps, simulate bottom-up attention by calculating saliency values, facilitating the detection of regions likely to capture attention [32].

In contrast, top-down attention is endogenous, driven by cognitive factors like expectations and task relevance. Top-down attention is voluntary and prioritizes specific stimuli based on cognitive goals [16] [7] [20]. It modulates sensory processing, directing attention toward goal-relevant stimuli while filtering out irrelevant information. While bottom-up attention focuses on sensory input saliency, top-down attention operates at a higher cognitive level, integrating contextual information and cognitive biases to guide attentional selection [22] [6]. Modeling top-down attention is challenging due to its complexity and reliance on internal cognitive states, making it an ongoing research focus in cognitive neuroscience and computational modeling.

	RGB Data	Event Data
Bottom-up	Itti et al. (1998) [13] Martinet et al. (2009) [21] Wang et al. (2010) [32] Gregor et al. (2015) [9]	Li et al. (2019) [18] Iacono et al. (2019) [11] Yao et al. (2021) [33] Liang et al. (2021) [19] Sun et al. (2022) [28] Gruel et al. (2022) [10] Bulzomi et al. (2023) [3]
Top-down	Zheng et al. (2015) [34] Radford et al. (2021) [24]	Cancini et al. (2019) [4] Kong et al. (2024) [15]

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

Table 1 offers a summarized overview of the bottom-up and top-down attention models that will be discussed in detail throughout the paper.

2.3. Top-down and bottom-up for RGB data

In this subsection, we will briefly review top-down and bottom-up models for RGB data before shifting more deeply into those for event data.

In 1998, Itti, Koch, and Niebur [13] introduce a computational model of visual attention that mimics the early stages of visual processing in primates by creating a saliency map through integrating various visual features at different spatial scales. The saliency map is then used to direct attention towards these regions, facilitating rapid analysis of the scene. This bottom-up approach depends solely on visual stimuli characteristics to direct attention, effective for initial scene analysis.

The study by Martinet et al. (2009) [21] explored gaze analysis as a tool for evaluating visual media quality by tracking viewers' eye movements to identify naturally attention-grabbing elements, emphasizing a data-driven, bottom-up approach based on the media's visual characteristics.

Wang et al. (2010) [32] examines experiments showing a strong relationship between visual salience (bottom-up) and visual importance (top-down) in the first two seconds of viewing. The findings reveal a moderate correlation between importance and saliency maps, which highlight shape and color, with early focus on human and animal faces. Importance maps prioritize categories like human and animal faces, influenced by the image's artistic meaning.

The Deep Recurrent Attentive Writer (DRAW) network by Gregor et al. (2015) [9] for image generation features a spatial attention mechanism inspired by human eye foveation. The model contains recurrent networks and dynamically updated attention, which iteratively build images through sequential modifications and create patches with varying locations and zoom levels. This bottom-up attention allows DRAW to adapt to different object scales and positions in complex visual tasks.

Zheng et al. (2015) [34] developed the Fixation NADE model, which uses visual attention by directing the recognition process through a sequence of task-specific objectives. Fixation NADE simulates human visual attention by learning what features to extract and where to extract them, using a fixation policy. This approach combines bottom-up and top-down attention. Key contributions include its autoregressive architecture for fixation-based recognition and its improved performance compared to earlier models like Fixation RBM.

CLIP (Contrastive Language-Image Pre-Training) model by Radford et al. (2021) [24] harnesses the rich information in natural language to enable zero-shot transfer learning, adapting to various tasks without additional labeled data. The architecture includes an image and a text encoder that identify correct image-text pairs. The model leverages top-down attention by using textual descriptions to guide focus on relevant image features. This innovative approach highlights the potential of natural language supervision in developing versatile visual models.

2.4. Top-down and bottom-up for event data

In contrast to RGB data, attention studies focusing on event data have been less prevalent. While there exists some studies of bottom-up, top-down models for event data remain relatively novel, with limited research due to their complexity and reliance on predefined tasks.

Li's 2019 paper [18] combines event data with frame-based images to enhance object detection for autonomous vehicles. The model employs two distinct processing streams: CNNs for frame-based data and SNNs for event data, generating visual attention maps. Integrating event cameras' high temporal resolution with traditional cameras' spatial detail, the approach improves detection accuracy in scenarios with rapid motion and varying lighting. Fusion of outputs from both streams is managed using Dempster-Shafer theory [17], resulting in improved performance on the DDD17 dataset [2]. This bottom-up approach enhances the model's ability to detect vehicles in challenging environments by combining neuromorphic and conventional vision technologies.

Iacono et al. (2019) [11] modified a proto-object attention model for neuromorphic event cameras to enhance iCub humanoid robot's visual processing capabilities. Using a bottom-up attention mechanism, the model processes event data through layers: center-surround filtering, border ownership cells, and grouping cells. These layers decompose the visual scene into proto-objects, with saliency determined by contrast and edges. The approach is validated with static and moving objects, showing higher object speeds lead to increased event counts, enhancing saliency. This efficient, low-latency method improves robotic vision in dynamic settings, ensuring the robot focuses on relevant stimuli.

Yao et al.'s 2021 research [33] 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. TA-SNN (Temporal-wise Attention Spiking Neural Networks) selectively filters non-essential event streams during inference based on their importance, optimizing resource usage. The model operates primarily with a bottom-up attention mechanism, processing event data by dynamically focusing on significant temporal segments determined by the data. TA-SNN achieved state-of-the-art results in classification tasks like gesture recognition, image classification, and spoken digit recognition, demonstrating effectiveness in handling sparse event streams.

Liang et al.'s 2021 paper [19] introduced a novel approach for event-based object detection using a lightweight spatial attention mechanism. The goal is to improve detection accuracy and efficiency in event data by minimizing noise and enhancing multi-scale feature maps through shallow feature integration. The method encodes event data into maps using Surface of Active Events (SAE) and Histogram of Intensities (HIS), followed by a Canny edge detector to highlight relevant features. A key innovation is the Spatial Attention Module (SAM), which integrates multi-scale spatial features with object detection. SAM focuses on areas with significant event activity, enhancing moving objects' detection by concentrating resources on RoI. The model uses a bottom-up approach, applying spatial attention to highlight significant features and improve detection accuracy in real-time automotive scenarios.

In autonomous driving, attention using event data is crucial for improving system capabilities in dynamic environments. Sun's 2022 research [28] introduces the Event Fusion Network (EFNet) designed to mitigate motion blur in images from conventional cameras. EFNet employs symmetric cumulative event representation and integrates Event-Image Cross-modal Attention (EICA) to blend data from event and frame-based cameras. The model operates in a two-stage architecture inspired by the event-based deblurring physical model [23], using bottom-up attention to highlight pertinent event stream features for accurate deblurring. Additionally, the paper introduces the RE-

Blur dataset, featuring real-world event streams paired with blurry and sharp images, to assess EFNet's performance. This study advances motion deblurring by utilizing event data with a bottom-up approach.

Gesture recognition is another application where attention using event data is highly effective. Gruel's 2022 study [10] demonstrates the advantages of neuromorphic attention for event data in recognizing hand and arm movements. Gruel's method involves a spiking neural network that dynamically adapts to incoming event streams, focusing on regions of high activity corresponding to meaningful areas while filtering irrelevant background noise. Regions with the highest event data density are defined as focal points of attention (RoI) within the model, enabling the system to prioritize significant activity while discarding irrelevant locations.

In object recognition from a static camera, attention using event data offers advantages, especially in dynamic environments and varying lighting conditions. Bulzomi's 2023 study [3] explores pedestrian detection using a static camera setup equipped with an event-based sensor. This method leverages the asynchronous nature of event data to detect scene changes, allowing the system to efficiently isolate and recognize moving objects against a static background. The definition of attention in Bulzomi's study aligns with Gruel's work [10], where RoI is determined as the position containing the highest density of events, ensuring the system focuses on the most pertinent data.

Cancini's 2019 research [4] presents a novel approach for object recognition using event cameras, focusing on enhancing processing efficiency in dynamic environments. The authors propose two models: the first 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 adapts the DRAW-based neural architecture [9] to handle event data effectively. These RoIs are processed by a Phased LSTM recognition network, combining spatial and temporal information for accurate object classification. The significant contribution of this research lies in adapting attention mechanisms for event cameras, demonstrating improved handling of translation and scale variations compared to conventional models.

Kong et al. (2024) [15] introduce a model for event-based semantic segmentation, leveraging the pre-trained knowledge from image and text domains using CLIP [24]. The objective is to address the challenges of dense annotations and scalability in event-based vision tasks by transferring semantically rich CLIP knowledge to event streams. The model employs a top-down approach, using text prompts to guide 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.

Amongst the state-of-the-art papers described in this 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

Our study aims to integrate both bottom-up and top-down attention mechanisms for event data, particularly in the context of autonomous driving. In this section, we will discuss the use of YOLO and YOLO-World as the bottom-up and top-down attention models, respectively, applied to the DSEC-Detection dataset. We will also present the data preprocessing steps, model fine-tuning process, and evaluation metrics used in the study.

3.1. YOLO – used as a bottom-up attention model

YOLO (You Only Look Once) [26] revolutionized object detection by consolidating the process into a single-stage framework. Unlike other models that use multiple processing stages, YOLO treats object detection as a direct regression task. This design enables the model to analyze the entire image in one pass using a convolutional neural network (CNN), allowing for fast and efficient real-time detection. YOLO architecture divides the input image into a grid, with each cell predicting bounding boxes, confidence scores, and class probabilities. Influenced by GoogLeNet, YOLO's architecture includes 24 convolutional layers and fully connected layers, with the Intersection over Union (IoU) metric ensuring precise object localization. YOLOv8, the latest version, further enhances performance, making it ideal for modern applications like autonomous vehicles, robotics, and real-time surveillance.

For bottom-up attention, we will use YOLOv8 due to its ability to learn from data without requiring predefined instructions. Known for its speed and accuracy, YOLOv8 will be fine-tuned on event images to assess its performance with this unique data type, as it was originally trained on frame-based datasets like COCO.

3.2. YOLO-World – used as a top-down attention model

YOLO-World [5] advances object detection by overcoming the limitations of traditional YOLO models, which are restricted to predefined categories. Built on the YOLOv8 architecture, YOLO-World integrates vision-language modeling through a Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN). This innovation merges visual features with text embeddings, enabling the detection of objects beyond fixed categories. A pre-trained text encoder (CLIP) converts text into embeddings that interact with visual features, enhancing the

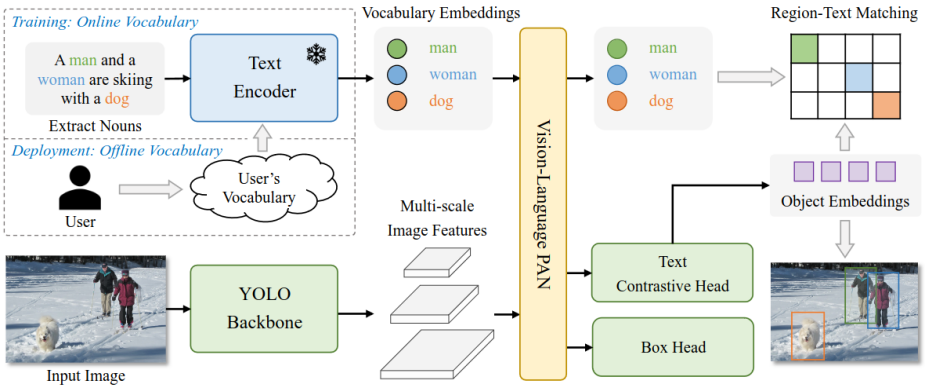


Figure 1. YOLO-World Architecture. Source: [5]

model's ability to match objects with textual descriptions. Figure 1 illustrates the YOLO-World architecture.

The architecture includes a backbone for feature extraction, a path aggregation network for multi-scale feature refinement, and a detection head for bounding box regression and object classification. A region-text contrastive loss aligns image regions with provided text, enabling zero-shot detection of objects not in the training set, guided by textual cues during inference. Additionally, an Intersection over Union (IoU) loss fine-tunes bounding box accuracy for precise localization. These features make YOLO-World a robust solution for open-vocabulary detection, ideal for dynamic real-world applications.

For top-down attention, YOLO-World [5] is selected due to its ability to specify detection classes before training, making it valuable for tasks requiring predefined object detection. This is especially useful in autonomous driving, where detection priorities may shift depending on the environment, such as focusing on "car" detection on highways or "pedestrian" detection in city centers. YOLO-World's ability to predict new classes not included in the original training set offers greater flexibility and adaptability in real-world scenarios. Like YOLO, YOLO-World requires fine-tuning on event data for optimal performance.

3.3. Models finetuning and evaluation metrics

YOLO and YOLO-World have been trained using RGB data so it is necessary to fine-tune them using event data. During the fine-tuning process, various combinations of hyperparameters—such as epoch count, batch size, and image size—were tested to identify the most effective configuration. The evaluation of YOLO and YOLO-World models focuses on key metrics: the confusion matrix, precision, recall, mAP50, mAP50-95, and various confidence curves, all of which help assess model accuracy and performance. Additionally, power consumption during inference is calculated, emphasizing the importance of energy efficiency for real-world applications, particularly in scenarios requiring low-energy consumption for each image processed. The detailed evaluation of these results, including performance across all classes, will be discussed in Section 4.

3.4. Dataset

DSEC-Detection dataset: The objective of this study is to work with event data, specifically focusing on object detection for autonomous driving scenarios. The dataset DSEC (Stereo Event Camera Dataset for driving scenarios) [8]

Data preprocessing The DSEC-Detection dataset provides event data. For this study, we use these event data to generate event images and corresponding labels, which are crucial for fine-tuning the attention models. The data processing involves several key steps: *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. *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. *Missing Value Removal:* Any missing values in the training, validation, and test datasets are removed to ensure data integrity. *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. For class distribution, 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.

4. Results and discussion

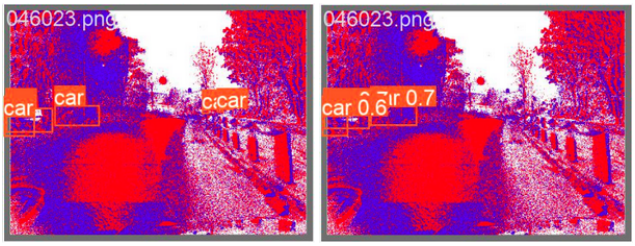


Figure 2. Illustration of YOLO detection of DSEC-Detection

We present in this Section results of YOLO and YOLO-World, followed by a detailed discussion of the findings.

4.1. Results obtained by YOLO

Results: To provide more insight into the detection outcomes, Figure 2 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, improved its ability to recognize objects in the new data format. As shown in Figure 2, YOLO managed to detect the "car" in event data that are challenging for the human eye to identify. This is a particularly noteworthy outcome of the experiment.

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. The Figure 3a shows that YOLO 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. The details for these metrics are presented in Table 2.

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

Table 2. YOLO metrics summary

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.

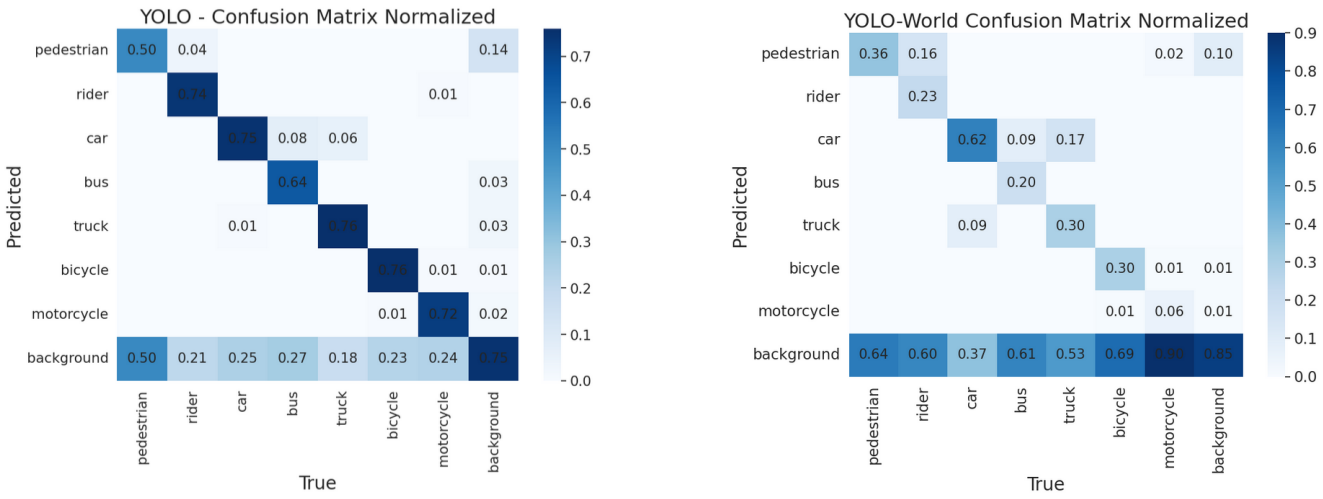
Moreover, the relatively low number of instances for other classes like "truck," "bicycle," and "motorbike" further limits the model's ability to learn and accurately predict these categories. The insufficient representation of these classes in the dataset likely contributes to the reduced effectiveness 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.

4.2. Results obtained by YOLO-World

Results: The primary focus of this study is the application of top-down attention for event data. In the example provided, the same moment is captured in an RGB image (Figure 4a), the corresponding event image (Figure 4b), and the predictions made by the fine-tuned YOLO-World model (Figure 4c). Remarkably, in Figure 4c, the model predicts a new class, "person," even though this class was not part of DSEC-Detection dataset. 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.

Besides the standard metrics (Precision, Recall, mAP50 and mAP50-95), we also considered the time and the power consumption of each model in order to find the optimal combination of hyper parameters. 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.

Tables 3 reveals 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. This configuration also requires the longest training time, which is expected since increasing the image size and the number of epochs



(a) Confusion matrix normalized - YOLO for DSEC-Detection

(b) Confusion matrix normalized - YOLO-World for DSEC-Detection

Figure 3. Comparison between confusion matrices of YOLO and YOLO-World for DSEC-Detection.

(a) Frame-based image

(b) Event image

(c) Prediction of YOLO-World

Figure 4. Comparison between frame-based image, event image, and YOLO-World prediction.

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. After considering the evaluation metrics and power consumption, the optimal results were achieved 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 4. These metrics were even more pronounced for the dominant "car" class, reflecting the model's stronger performance when sufficient data is available for training.

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 precision, recall, mAP50, and mAP50-95 scores of the YOLO-World optimal model are higher than those achieved with YOLO. Besides, it is important to note that the accuracy for all classes detected by YOLO-World is lower in comparison to the detection by YOLO, with 0.62 of accuracy for "car". The other less dominant classes such as "truck" and "bicycle" receive 0.3 accuracy. The detailed results of accuracy for YOLO-World is presented in Figure 3b.

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

Image size	epoch	batch	P	R	mAP50	mAP50-95	Time train (hour)	Time total (hour)	Reference time per image (ms)	Power consumption per image (J)
256	7	32	0.442	0.214	0.220	0.132	3.515	3.866	14.3	1.1154
512	7	32	0.596	0.343	0.371	0.239	10.760	11.273	98.3	7.469
512	20	32	0.594	0.342	0.374	0.248	27.306	27.521	27.0	2.025

Table 3. YOLO-World Metrics – Comparison with different combinations of hyperparameters, including Power Consumption Metrics

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

Table 4. YOLO-World Metrics Summary with optimal combination of hyperparameters

to learn and predict more effectively. 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: *Model Architecture:* YOLO is optimized for speed and may sacrifice some accuracy, especially in object localization. YOLO-World, with its multimodal approach, uses text prompts to improve accuracy in distinguishing between objects and background, leading to better precision, recall, and mAP. *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. *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.

5. Conclusion

Event cameras, which capture brightness changes, present various advantages such as high temporal resolution and low latency, are ideal for dynamic environments. This paper explores cognitive attention including bottom-up attention, based on sensory input, and top-down attention, guided by task relevance and prior knowledge. Despite the potential of cognitive attention mechanisms in event data for computer vision applications, there is a lack of research in this area, particularly top-down attention for event data.

Our paper presents three major contributions. First, we address the existing gap in understanding and terminology related to cognitive attention mechanisms for event data by providing a comprehensive survey, which is a valuable resource for future research. Second, we review and categorize existing studies that employ cognitive attention models for both RGB and event data, focusing on various computer vision applications. These studies are classified into either top-down or bottom-up attention models, enhancing clarity in the field. Finally, we demonstrate the application of YOLO and YOLO-World on the DSEC-Detection dataset. By testing bottom-up and top-down attention mechanisms on a large event dataset for autonomous driving tasks, our experiments yield promising results that indicate strong potential for future research and practical implementations.

The research also highlights challenges. Model performance was affected by class distribution, with dominant class like "car" scoring higher in accuracy than less frequent ones. Distinguishing between background and object classes in event images was difficult, lowering some performance metrics. Future researches could consider more balanced datasets or advanced data augmentation techniques. Moreover, to improve the performance of event-based vision systems, we suggests using manually precise, and detailed descriptions or prompts instead of automatically generated ones. These enhancements are expected to significantly advance the performance and reliability of attention models for event data.

In conclusion, the exploration of cognitive attention mechanisms in event data is promising for advancing modern vision systems. This paper lays a foundation for future research, in survey and applications, emphasizing the importance of integrating these mechanisms to create more efficient and accurate vision systems.

Acknowledgments

This work was supported by (HIDDEN FOR BLIND REVIEW).

References

[1] DSEC-Detection. Online. https://dsec.ifi.uzh.ch/dsec-detection/. 1

[2] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017. 3

[3] Hugo Bulzomi, Amélie Gruel, Jean Martinet, Takeshi Fujita, Yuta Nakano, and Rémy Bendahan. 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)*, pages 1–5. IEEE, 2023. 2, 4

[4] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Attention mechanisms for object recognition with event-based cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019. 2, 4

[5] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 1, 4, 5

[6] Charles E Connor, Howard E Egeth, and Steven Yantis. Visual attention: bottom-up versus top-down. *Current biology*, 14(19), 2004. 2

[7] Simone Frintrop, Erich Rome, and Henrik I Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):1–39, 2010. 2

[8] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3), 2021. 5

[9] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015. 2, 3, 4

[10] Amélie Gruel, Antonio Vitale, Jean Martinet, and Michele Magno. Neuromorphic event-based spatio-temporal attention using adaptive mechanisms. In *2022 IEEE 4th international conference on artificial intelligence circuits and systems (AICAS)*. IEEE, 2022. 2, 4

[11] Massimiliano Iacono, Giulia D'Angelo, Arren Glover, Vadim Tikhanoff, Ernst Niebur, and Chiara Bartolozzi. Proto-object based saliency for event-driven cameras. In *IROS*. IEEE, 2019. 2, 3

[12] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3), 2001. 2

[13] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 2

[14] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 1

[15] Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R Cottereau, and Wei Tsang Ooi. Openess: Event-based semantic scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 4

[16] Patrick Le Callet and Ernst Niebur. Visual attention and applications in multimedia technologies. *Proceedings of the IEEE*, 101(9):2058–2067, 2013. 2

[17] Hyungtae Lee, Heesung Kwon, Ryan M Robinson, William D Nothwang, and Amar M Marathe. Dynamic belief fusion for object detection. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016. 3

[18] Jianing Li, Siwei Dong, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Event-based vision enhanced: A joint detection framework in autonomous driving. In *2019 ieee international conference on multimedia and expo (icme)*. IEEE, 2019. 2, 3

[19] Zichen Liang, Guang Chen, Zhijun Li, Peigen Liu, and Alois Knoll. Event-based object detection with lightweight spatial attention mechanism. In *2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2021. 2, 3

[20] Grace W Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14, 2020. 2

[21] Jean Martinet, Adel Lablack, Stanislas Lew, and Chabane Djeraba. 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. 2

[22] Behrad Noudoost, Mindy H Chang, Nicholas A Steinmetz, and Tirin Moore. Top-down control of visual attention. *Current opinion in neurobiology*, 20(2), 2010. 2

[23] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. 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. 3

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 4

[25] Yashas Rai, Patrick Le Callet, and Gene Cheung. 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. 2

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 4

[27] Davide Scaramuzza. Tutorial on event-based cameras. In *IROS 2015: Proc. of the 2nd Workshop on Alternative Sensing for Robot Perception*, 2015. 1

[28] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European conference on computer vision*. Springer, 2022. 2, 3

[29] John K Tsotsos. *A computational perspective on visual attention*. MIT Press, 2021. 2

[30] Mohit Vaishnav. *Exploring the role of (self-) attention in cognitive and computer vision architecture*. PhD thesis, Université Paul Sabatier-Toulouse III, 2023. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[32] Junle Wang, Damon M Chandler, and Patrick Le Callet. Quantifying the relationship between visual salience and visual importance. In *Human vision and electronic imaging XV*, volume 7527. SPIE, 2010. 2

[33] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-wise attention spiking neural networks for event streams classification. In *CVPR*, 2021. 2, 3

[34] Yin Zheng, Richard S Zemel, Yu-Jin Zhang, and Hugo Larochelle. A neural autoregressive approach to attention-based recognition. *IJCV*, 113, 2015. 2, 3