Depth and Event-Based Approaches for Human Detection and Pose Estimation

Hao Quan[1(✉)], Milad Goudarzi[1], Chiara Plizzari[1], Simone Mentasti[1], Francesca Palermo[2], Diana Trojaniello[2], and Matteo Matteucci[1]

[1] Politecnico di Milano, Milan, Italy
hao.quan@polimi.it
[2] EssilorLuxottica, Milan, Italy

Abstract. RGB cameras, while widely used for human detection and pose estimation, face challenges in real-world applications due to sensitivity to illumination changes, low-light conditions, and motion blur. Moreover, deploying state-of-the-art models is hindered by privacy concerns and the high computational demands of RGB images, making them less suitable for real-time, resource-constrained environments. This paper investigates depth and event-based cameras as alternatives to RGB for human detection and pose estimation. These modalities offer advantages such as improved low-light performance and reduced privacy concerns. Event cameras, which capture pixel-level intensity changes at high temporal resolution, minimize motion blur and consume significantly less power and memory than RGB cameras, making them ideal for edge applications. We simulate depth and event-based data using the MS COCO dataset and retrain state-of-the-art models. Results show depth data performs similarly to RGB, while event-based data, though slightly less accurate, remains competitive. These findings highlight the potential of depth and event-based cameras for efficient, privacy-preserving human detection and pose estimation in real-world applications.

Keywords: Human detection · Human Pose estimation · Event-based Cameras · Depth

1 Introduction

Human detection and pose estimation are critical computer vision tasks that enable technologies to understand and interact with the physical world. These tasks have diverse applications, including autonomous vehicles [26,29] and remote sensing systems [19,31]. By observing human position, motion, and actions in the surroundings, these technologies enhance user awareness of their environment, improving overall user experience.

Over the past decade, deep learning-based models, particularly the YOLO (You Only Look Once) [22] and YOLO-Pose [14] families, have gained popularity for their speed and accuracy in real-time human detection and pose estimation.

E. Rodolà et al. (Eds.): ICIAP 2025, LNCS 16167, pp. 207–219, 2026.
https://doi.org/10.1007/978-3-032-10185-3_17

However, state-of-the-art person detection and pose estimation models face significant challenges when deployed in real-world scenarios, such as lighting variations, complex backgrounds, occlusions, particularly on edge devices like smart eyewear [18]. Wearable devices are often subject to strong variations in lighting conditions, including low-light environments [13], which can hinder the effectiveness of RGB-based models. Alternative data sources, including depth sensors and event-based cameras [6], present promising solutions to these challenges. Figure 1 provides a comparison of the RGB, depth, and event-based modalities for the same scene. These modalities offer distinct advantages, including improved performance in low-light and complex background conditions, lower power consumption compared to RGB cameras, and a reduced risk of revealing sensitive visual details, making them more suitable for privacy persevering applications. However, research in this domain remains limited, with notable gaps in publicly available datasets, optimized methodologies, and practical deployment strategies.

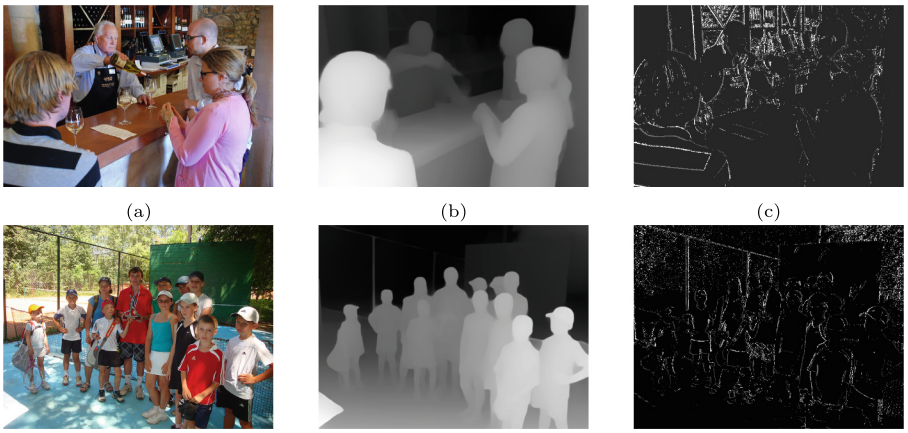

Fig. 1. Visualization of the same scene from the COCO dataset [11] captured in three modalities: (a)(d) RGB provides a standard visual representation, (b)(e) depth encodes object distances from the camera, and (c)(f) the event-based modality detects brightness changes over time, enhancing dynamic scene perception.

This paper investigates how these unconventional modalities can achieve performance comparable to RGB-based approaches for human detection and pose estimation, while offering additional benefits such as lower power consumption, robustness to complex visual conditions and enhanced privacy. We used the large-scale MS COCO object detection dataset [11] as our benchmark. To simulate the depth and event-based data modalities, we employed Depth Anything [30] and the Prophesee Metavision SDK [21], respectively. We then re-trained YOLO and YOLO-Pose 5 on various modalities of the COCO dataset, including RGB,

depth, and event-based data, for the human detection and human pose estimation tasks respectively. Results show that the performance of the depth-based modality is on-par with that of the traditional RGB modality, while the performance of the event-based modality is slightly lower than RGB. However, event-based cameras still provided competitive results, all while offering the advantages, such as robustness to illumination changes and complex backgrounds and lower power consumption.

The key contributions of this work are:

- We simulated event-based and depth data from the RGB images in the MS COCO dataset and released it to the research community, facilitating further advancements in human detection and human pose estimation with multiple modalities.
- We benchmarked YOLO / YOLO-Pose 5 human detection and pose estimation algorithms across different image modalities, including RGB, depth, and event-based data using the MS COCO dataset.
- Our experiments demonstrated that re-trained YOLO and YOLO-Pose models perform well on depth and event-based data from the simulated MS COCO datasets, making them promising for future work in human detection and human pose estimation tasks.

2 Related Works

2.1 Human Detection and Pose Estimation

Human Detection. Human detection has seen considerable advancements with various deep learning-based methodologies. Generally, these techniques can be categorized into two main approaches: region proposal-based two-stage methods and one-stage methods. Two-stage methods, such as R-CNN [7] and Faster R-CNN [23], first generate region proposals and then classify them, getting high detection accuracy at the cost of increased computational complexity. On the other hand, YOLO (You Only Look Once) [22] algorithm is a popular one-stage detection framework that integrates region proposal and classification into a single step, prioritizing speed and simplicity. Other notable one-stage method include Single-Shot Refinement Neural Network (RefineDet) [34], which improves detection accuracy by refining predictions. Transformer-based approaches such as DETR [3] use self-attention for solving detection tasks in an end-to-end manner, directly predicting bounding boxes by modelling relatinships between image regions without need for hand-crafted anchors or region proposals.

Human Pose Estimation. Human pose estimation algorithms divide into 2D and 3D approaches. However, the limited computing resources of edge devices make 2D methods more suitable for edge devices. Among those, OpenPose [2] and HRNet [4] have established benchmarks in the field by using multistage architectures and high-resolution feature maps to predict human keypoints with

high accuracy. YOLO-Pose [14] extends the capabilities of the YOLO object detection framework by integrating pose estimation into a unified architecture, enabling simultaneous detection and pose estimation.

In this paper, we use YOLO and YOLO-Pose due to their efficiency and adaptability, which make them well-suited for real-time applications and deployment on edge devices. Furthermore, we leverage these algorithms to explore the potential of human detection and pose estimation on alternative data modalities beyond RGB, such as depth and event-based data.

2.2 Event-Based Human Detection and Pose Estimation Tasks

While conventional human detection is well-established in computer vision, *event-based human detection* remains in its early stages. It holds great promise for real-time applications due to the low latency and power efficiency of event cameras, but also poses new challenges due to the shift from frame-based to event-based data. Morever, due to the challenges in data collection and annotation, the availability of annotated event-based datasets is significantly limited compared to other common modalities, limiting the research. PEDRo [1] offers a large-scale event-based human detection dataset from mobile robots, which was collected from public spaces with different lighting and meteorological conditions. eTraM [28] provides a novel event-based dataset for traffic monitoring that includes human detection annotations. Besides the event-based modality, MMPedestron [35] provides multi-modal data sources (RGB, IR, Depth, and LiDAR) for pedestrian detection. All the above approaches explore event-based human detection by converting event streams into frame-like representations followed by standard detection algorithms.

Recent works have adapted state-of-the-art *pose estimation models for event-based data*. EventHPE [36] is a two-stage deep learning framework for event-based 3D human pose and shape estimation. The first FlowNet stage infers optical flow from events, while the second ShapeNet stage estimates 3D shapes using both event data and flow. Yu et al. [32] proposed an adaptive vision transformer with adaptive patch sampling and adaptive token reduction, optimizing efficiency while maintaining accuracy. They also provided a large-scale event-based human pose estimation dataset EventMM HPE.

While event-based human detection and pose estimation have typically been treated as separate tasks, we propose a unified framework that jointly tackles both, leveraging not only RGB, but also depth and event-based data for robust performance in diverse real-world scenarios. To support this effort, we introduce a large-scale dataset specifically designed to enable the joint study of event-based human detection and pose estimation. Unlike prior works that focus mainly on pedestrian detection in constrained or driving environments, our dataset captures a broad range of contexts, covering varied activities and settings.

3 Event-Based Vision

Event-Based Data. Unlike traditional cameras that capture frames at fixed intervals, event cameras asynchronously record brightness changes at each pixel. Pixels in event cameras [6] operate independently and respond to variations in the continuous log-brightness function $L(\mathbf{u}, t)$, where $\mathbf{u} = (x_k, y_k)^T$ represents the spatial coordinates of a pixel, and t denotes the timestamp at which the brightness change occurs. An event is represented as a tuple $e_k = (x_k, y_k, t_k, p_k)$, where (x_k, y_k) denotes the pixel location, t_k is the timestamp, and $p_k \in \{-1, 1\}$ represents the polarity of the brightness change (indicating increase or decrease). An event is triggered whenever the change in log-brightness at pixel $\mathbf{u} = (x_k, y_k)^T$ exceeds a predefined threshold C, given that a sufficient time interval has elapsed since the last event at the same pixel. This is mathematically formulated as:

$$\Delta L(\mathbf{u}, t_k) = L(\mathbf{u}, t_k) - L(\mathbf{u}, t_k - \Delta t_k) \geq p_k C. \tag{1}$$

Thus, the output of an event-based camera consists of a continuous stream of events, represented as the sequence:

$$\mathcal{E} = \{(x_k, y_k, t_k, p_k) \mid t_k \in \tau\}, \tag{2}$$

where τ defines the time interval within which events occur.

Event-Based Representation. To generate a frame at a precise time t_k, events should be accumulated over a longer period of time (between the time $t_k - \Delta t_k$ and time t_k). Note that Δt_k is usually called accumulation time. To generate a frame with the accumulated events, for each event occurring between the time $t_k - \Delta t_k$ and time t_k, a white pixel is stored if the polarity of the event is positive, and a grey pixel if the polarity is negative. As it can be seen in Fig. 1 the background is set to black. This is the so-called *event counts* representation [15]. The resulting representation is an image of size $H \times W$.

4 Methodology

This work focuses on human detection and pose estimation from RGB, depth, and event-based images. In this section, we first define the tasks (Sect. 4.1) and describe the dataset used (Sect. 4.2), followed by an overview of the pipeline and models (Sect. 4.3). This section also includes a description of our approach to simulate depth and event-based data, as well as the evaluation metrics used.

4.1 Task Definition

Human Detection. Human detection aims to identify and localize individuals within a given input image. The model takes a two-dimensional image as input and predicts the regions containing humans by outputting bounding boxes defined by four coordinates $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$. Additionally, each detection is associated with a confidence score ranging from 0 to 1, indicating the likelihood of the detected region containing a person.

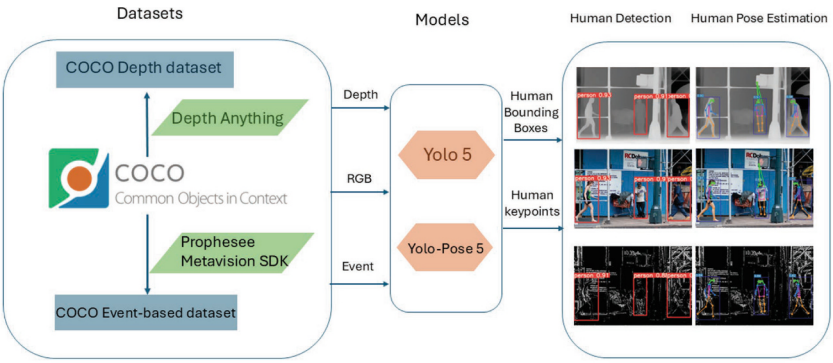

Fig. 2. Starting from the MS COCO RGB dataset, we simulate depth data generated using Depth Anything, and event data with the Prophesee Metavision SDK [21]. We then use YOLO 5 and YOLO-Pose 5 for producing human detection and pose estimation results across depth, RGB, and event-based modalities.

Human Pose Estimation. Human pose estimation involves predicting the 2D coordinates of 17 keypoints corresponding to human joints from images. Given an input image, the model estimates a predefined set of keypoints, such as shoulders, elbows, and knees, along with their respective confidence scores. The output consists of keypoint coordinates (x_i, y_i) for each detected individual, typically visualized as a skeletal structure.

4.2 Datasets

We use the MS COCO dataset [11] as the baseline for our experiments. It includes 118,287 training images, 5,000 validation images, and 40,670 test images, with a total of 164,957 images spanning 80 object categories. Each image contains multiple object instances, and for human-related tasks, the dataset provides 17 keypoint annotations per person, covering joints such as the head, shoulders, elbows, wrists, hips, knees, and ankles–enabling detailed pose estimation.

Building on this, we introduce depth-based and event-based versions of the dataset, empowering the community to explore human detection and pose estimation using multiple modalities.

4.3 Pipeline

Overview. The complete pipeline of the methodology is shown in Fig. 2. Specifically, starting from the original MS COCO RGB dataset, we introduced two simulated datasets: a MS COCO Depth dataset and a MS COCO Event-based dataset. These datasets enable a comprehensive evaluation of human detection and pose estimation under diverse visual inputs. We employ YOLO [22] and YOLO-Pose [14] in all experiments, as they offer a unified framework for both human detection and pose estimation–making them well-suited for our joint task.

We test these models across RGB, depth, and event-based modalities, demonstrating the adaptability of our approach to different data modalities.

Model Description. You Only Look Once (YOLO) [22] is a real-time object detection algorithm that performs object localization and classification in a single forward pass of a deep neural network. Unlike region-based approaches, YOLO formulates detection as a regression problem, dividing the input image into an $S \times S$ grid, where each cell predicts B bounding boxes, objectness scores, and class probabilities. The final detection is obtained by filtering predictions using non-maximum suppression [17]. This one-stage detection framework enables high-speed inference while maintaining competitive accuracy, making it suitable for applications. Given an input image I, YOLO predicts a set of bounding boxes $\mathcal{B}$, where each box b_i is defined as:

$$b_i = (x_i, y_i, w_i, h_i, c_i),$$

where (x_i, y_i) represents the center coordinates, (w_i, h_i) the width and height, and c_i the confidence score

YOLO-Pose [14] extends the YOLO detection framework to predict keypoints for human pose estimation. In addition to detecting objects, the model regresses multiple keypoint coordinates per detected entity, enabling real-time multi-person pose estimation. Given an input image I, YOLO Pose predicts a set of bounding boxes $\mathcal{B}$, where each detected entity b_i is defined as:

$$b_i = (x_i, y_i, w_i, h_i, c_i, K_i),$$

where (x_i, y_i) represents the bounding box center coordinates, (w_i, h_i) the width and height, c_i the confidence score, and $K_i = \{(k_{x1}, k_{y1}), \ldots, (k_{xN}, k_{yN})\}$ represents the set of N keypoints associated with the object. Each keypoint $k_j = (k_{xj}, k_{yj})$ corresponds to a specific body joint.

Model Selection. The selection of the specific model versions was based on specific criteria: (1) implementation in plain PyTorch without proprietary dependencies, excluding YOLO 9–11 [27]; (2) consistency between YOLO and YOLO-Pose to streamline modifications, ruling out YOLO 6 due to the lack of a publicly released YOLO-Pose 6 [10]; and (3) community validation, favoring YOLO/YOLO-Pose 5, which was released at CVPR 2022 [14] and includes optimized versions for edge devices. YOLO/YOLO-Pose 8 lacks similar validation. After evaluating available versions, YOLO/YOLO-Pose 5 was selected for its balance of precision, parameters, and model size, making it ideal for real-time human detection and pose estimation on edge devices.

Data Simulation. Since the original MS COCO dataset does not include depth or event-based modalities, we generated corresponding versions by utilizing Depth Anything [30] for depth simulation and the Prophesee library [21]

for event-based modality. We make both datasets publicly available to the community, with the aim of supporting further research and development in human detection and pose estimation using non-RGB modalities.

Metrics. For human detection, we use mean Average Precision (mAP) 0.5–0.95 as the evaluation metric. For human pose estimation, we also employed mean Average Precision (mAP) as the evaluation metric, but with a key difference in its computation. Instead of using bounding boxes, mAP for pose estimation is based on the alignment of predicted keypoints with ground truth keypoints, evaluated using the Object Keypoint Similarity (OKS) metric [16].

5 Experiments

In this section, we describe the experiments conducted to evaluate the proposed methods. We begin with implementation details (Sect. 5.1), followed by results across depth, event-based, and RGB modalities on the MS COCO dataset [11] (Sect. 5.2). Finally, we present some qualitative results (Sect. 5.3).

5.1 Implementation Details

The hyperparameters used in the experiments are consistent with the original YOLO 5 and Yolo-Pose 5 settings on the COCO RGB modality. Specifically, the models were trained using a batch size of 16 and SGD optimizer [24] with a one-cycle learning rate policy [25], where the learning rate follows a cosine decay schedule [12]. The initial learning rate was set to 0.01, reaching a peak of 0.002 before gradually decreasing throughout the training process.

All experiments were conducted using an NVIDIA RTX A6000 GPU. We used the models with best mAP during training for validation.

5.2 Results

Depth. Table 1 presents the performance on the human detection task of the YOLO 5 s, YOLO 5 m, and YOLO 5l models on both RGB data and the simulated depth version of the MS COCO dataset, as described in Sect. 4.3. For the COCO RGB dataset, the models achieved mAP scores of 66.22%, 69.15%, and 70.16%, respectively. When evaluated on the depth modality, the mAP scores across IoU thresholds from 0.5 to 0.95 were 64.3%, 66.28%, and 66.58%, respectively. Similarly, experiments on pose estimation task with YOLO-Pose 5 s, YOLO-Pose 5 m, and YOLO-Pose 5l were conducted on the simulated depth COCO dataset. The mAP scores at IoU thresholds from 0.5 to 0.95 were 61.29%, 62.1%, and 65.15%, compared to 63.44%, 67.06%, and 69.47% on the MS COCO RGB dataset for YOLO-Pose 5 s, YOLO-Pose 5 m, and YOLO-Pose 5l, respectively. For the smaller model YOLO 5 s, the performance gap between RGB and depth is only around 2%. As the model size increases, the gap becomes slightly

more pronounced. For example, in the case of YOLO 5l, the performance difference between RGB and depth increases to approximately 4%. These trends are consistent across YOLO-Pose models, with larger models also showing a slightly larger gap between modalities. These results indicate that while the performance of YOLO and YOLO-Pose models on the depth modality of the MS COCO dataset is slightly lower than their performance on the RGB modality, the differences are minimal, indicating comparable effectiveness across modalities.

Table 1. mAP results of re-training and evaluating human detection YOLO 5 and human pose estimation YOLO-Pose 5 on different modalities of the COCO dataset.

Model	RGB (%)	Depth (%)	Event (%)
Human Detection			
YOLO 5 s	66.22	64.30	44.45
YOLO 5 m	69.15	66.28	47.90
YOLO 5l	70.16	66.58	48.82
Human Pose Estimation			
YOLO-Pose 5 s	63.44	61.29	43.03
YOLO-Pose 5 m	67.06	62.10	47.30
YOLO-Pose 5l	69.47	65.15	49.14

Event. In Table 1 we also present results obtained on the simulated event-based data. We re-trained and evaluated YOLO 5 s, YOLO 5 m, and YOLO 5l on the simulated event-based MS COCO dataset, obtaining mAP scores 44.45%, 47.9%, and 48.82%, respectively. When compared to performance on the MS COCO RGB dataset, the models achieved mAP scores of 66.22%, 69.15%, and 70.16%, respectively. The mAP scores for YOLO-Pose 5 s, YOLO-Pose 5 m, and YOLO-Pose 5l were 43.03%, 47.3%, and 49.14%, respectively, which were slightly lower than the results on the RGB modality, where the scores were 63.44%, 67.06%, and 69.47%, respectively. Notably, the event modality has historically lagged significantly behind RGB in classification tasks. For instance, the recent N-ImageNet benchmark [9] reports an accuracy of 48.94% for the best-performing event-based architecture, markedly lower than the >90% accuracy achieved by RGB models [5,8,20,33]. Our results indicate that, although YOLO and YOLO-Pose perform worse on event-based data compared to RGB data, the performance gap is not substantial. As demonstrated in the qualitative results in the next section, event-based data remains a competitive modality for human detection and pose estimation tasks.

5.3 Qualitative Results

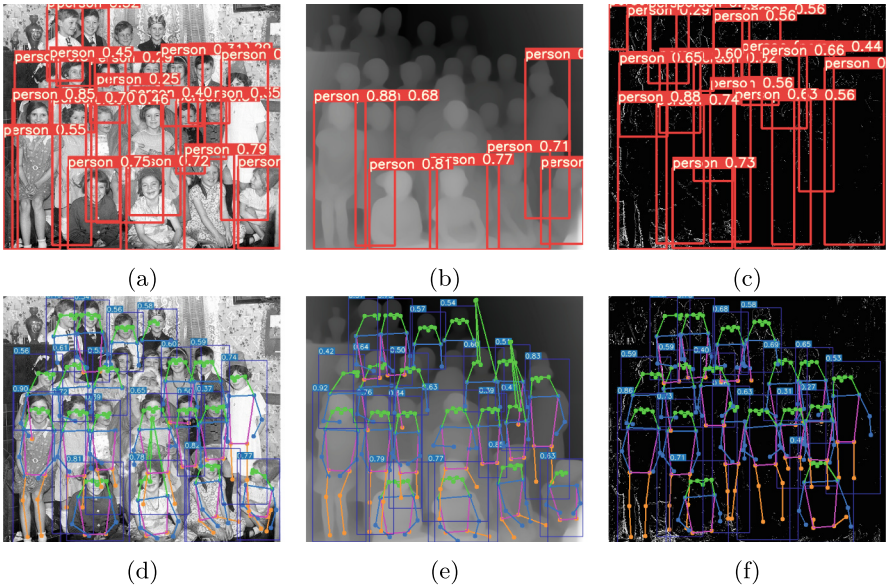

Fig. 3. Results from our re-trained YOLO 5 model for human detection (top row: a-c) and YOLO-Pose 5 for pose estimation (bottom row: d-f) on a MS COCO scene across RGB, depth, and event modalities. Red bounding boxes and scores show detected people. Blue boxes, green keypoints, and colored skeletons visualize estimated poses. (Color figure online)

For these analyses, we used our re-trained YOLO 5 and YOLO-Pose 5 models, which achieved the highest mAP scores on the RGB, depth, and event-based modalities of the MS COCO validation set (Sect. 5.2). These models were applied for human detection and pose estimation on the same MS COCO images represented in RGB (Fig. 3a, Fig. 3d), depth (Fig. 3b, Fig. 3e), and event-based (Fig. 3c, Fig. 3f) formats. The results show strong performance in RGB for both tasks. In the depth modality, accuracy remains high with few people, but in crowded scenes, pose estimation detects more individuals (Fig. 3e) than the detection model (Fig. 3b). However, YOLO-Pose may misassign head joints when people are close together, as seen in the upper right of Fig. 3e, highlighting challenges in resolving closely positioned individuals. While Table 1 indicates that event-based data yields lower mAP than RGB and depth, qualitative results remain competitive, especially as crowd size increases (Fig. 3c, Fig. 3f). These findings support depth and event-based modalities as effective alternatives for human detection and pose estimation.

6 Conclusion

In this work, we explored the potential of various data modalities for human detection and pose estimation. Traditional RGB data struggles in challenging lighting and complex backgrounds, so we investigated alternatives like depth and event-based data. Our experiments showed that re-trained YOLO and YOLO-Pose models perform well on non-RGB data modalities, including depth and event-based data, using the MS COCO dataset. We conclude that depth and event-based data can effectively handle lighting and background challenges, making them viable alternatives to RGB for these tasks.

Acknowledgements. This work was conducted at Smart Eyewear Lab, a joint research center between EssilorLuxottica and Politecnico di Milano.

References

1. Boretti, C., Bich, P., Pareschi, F., Prono, L., Rovatti, R., Setti, G.: PEDRo: an event-based dataset for person detection in robotics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4065–4070 (2023)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229. Springer (2020)
4. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: HigherHRNet: scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5386–5395 (2020)
5. Dai, Z., Liu, H., Le, Q.V., Tan, M.: CoAtNet: marrying convolution and attention for all data sizes. Adv. Neural. Inf. Process. Syst. **34**, 3965–3977 (2021)
6. Gallego, G., et al.: Event-based vision: a survey. TPAMI **44**(1), 154–180 (2020)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR, pp. 16000–16009 (2022)
9. Kim, J., Bae, J., Park, G., Zhang, D., Kim, Y.M.: N-ImageNet: towards robust, fine-grained object recognition with event cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2146–2156 (2021)
10. Li, C., et al.: YOLOv6: a single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022)
11. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: ECCV, pp. 740–755. Springer (2014)
12. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

13. Mai, C., et al.: DBCG-Net: dual branch calibration guided deep network for UAV images semantic segmentation. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. (2024)
14. Maji, D., Nagori, S., Mathew, M., Poddar, D.: YOLO-Pose: enhancing yolo for multi person pose estimation using object keypoint similarity loss. In: CVPR, pp. 2637–2646 (2022)
15. 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)
16. Microsoft: COCO keypoint detection challenge. https://cocodataset.org/#keypoints-2020 (2020). Accessed 20 Jan 2025
17. Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR'06), vol. 3, pp. 850–855. IEEE (2006)
18. Palermo, F., et al.: Advancements in context recognition for edge devices and smart eyewear: sensors and applications. IEEE Access (2025)
19. Pan, J., et al.: MSFA-Net: multiple spatial-channel feature aggregation network for change detection and a UAV-CD dataset. In: IGARSS, pp. 10328–10332. IEEE (2024)
20. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: CVPR, pp. 11557–11568 (2021)
21. Prophesee: https://www.prophesee.ai/ (2024)
22. Redmon, J.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
24. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat., 400–407 (1951)
25. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820 (2018)
26. Song, Z., et al.: Robustness-aware 3D object detection in autonomous driving: a review and outlook. TPAMI (2024)
27. Ultralytics: YOLO. https://www.ultralytics.com/ (2025). Accessed 12 Jan 2025
28. Verma, A.A., Chakravarthi, B., Vaghela, A., Wei, H., Yang, Y.: eTraM: event-based traffic monitoring dataset. In: CVPR, pp. 22637–22646 (2024)
29. Xian, T., et al.: A scale-temporal interaction network for remote sensing image change detection and a UAV-CD dataset. In: IGARSS, pp. 8603–8607. IEEE (2024)
30. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: unleashing the power of large-scale unlabeled data. In: CVPR, pp. 10371–10381 (2024)
31. Ying, Z., et al.: Large-scale high-altitude UAV-based vehicle detection via pyramid dual pooling attention path aggregation network. IEEE Trans. Intell. Transp. Syst. (2024)
32. Yu, N., et al.: Adaptive vision transformer for event-based human pose estimation. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 2833–2841 (2024)
33. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. arXiv preprint arXiv:2106.04560 (2021)

34. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: CVPR, pp. 4203–4212 (2018)
35. Zhang, Y., Zeng, W., Jin, S., Qian, C., Luo, P., Liu, W.: When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset. In: ECCV, pp. 430–448. Springer (2025)
36. Zou, S., et al.: EventHPE: event-based 3D human pose and shape estimation. In: CVPR, pp. 10996–11005 (2021)