Assessment of deep learning-based detection algorithms using event cameras for construction applications

Robert Guamán-Rivera, Ariel Zuniga-Santana, and Rodrigo Verschae
Institute of Engineering Sciences, Universidad de O'Higgins, Chile

***Abstract*— Collaborative work of robots and workers faces several challenges during 3D printing process in the construction industry. In this regard, implementing deep learning-based detection algorithms represents a promising technology to improve construction tasks in dynamic robot-worker interaction environments in 3D printing applications. However, the visual perception of the environment with conventional cameras has the disadvantage of being susceptible to changes in lighting and dynamic changes in the construction process. In this context, using technology based on bio-inspired sensors could overcome the critical problems conventional cameras face, considering that its characteristics are high temporal resolution, high dynamic range, low power consumption and high bandwidth. This paper compares two convolutional neural network-based object detection algorithms designed to identify workers, hard hats, trees, tables and robotic arms within an environment emulating a virtual construction site. To this end, the You Only Look Once (YOLO) v8 and Detection Transformer (DETER) algorithms have been trained and experimentally tested using various event frames in a dynamic emulated construction environment. Experimental results revealed that the DETER algorithm exhibits a higher detection performance, achieving an precision rate of 92.1%.**

I. INTRODUCTION

Currently, the construction industry is implementing digital BIM technologies (Building Information Model), virtual reality (VR) and artificial intelligence (AI) algorithms in additive manufacturing processes with 3D printing. These technological advances are crucial to enhancing conventional construction's efficiency, accuracy and sustainability [1].

Manual tasks in the construction sector are emerging towards collaboration between workers and robots to improve productivity and safety by combining robotic intelligence with workers' skills. Compared to pre-programmed autonomous robots, collaborative robotics emphasises worker competencies, such as intuitive decision-making, responsiveness and adaptability. At the same time, dynamic programming and perception of robots become crucial to executing complex tasks in a collaborative construction environment [2].

In addition, construction processes will benefit from integrating robotic platforms equipped with sensors employing AI algorithms and simulation software in the construction sector [1]. Since construction processes are dynamic, integrating robots offers significant benefits by ensuring the efficient and accurate execution of complex tasks [3].

The interaction between robotic platforms on a construction site, and workers and their work in a shared space, is one of the main challenges in the construction sector. In this context, AI algorithms and computer vision allow the development of detection strategies for workers, fixed and moving objects and safety equipment to ensure the correct monitoring of the active agents involved in the construction process [1], [4].

However, the challenges of implementing AI algorithms are associated with the input data quality and the sensors' technical specifications used to acquire data from the environment [4]. Whereas the use of RGB cameras in monitoring processes in construction applications has increased, these devices still present certain disadvantages of blurring and high sensitivity to environmental and hardware changes when implementing AI algorithms [5], [6].

In this paper, we explore a novel technology known as an event camera or neuromorphic sensor; this sensor registers the intensity changes of each pixel within its field of vision and provides relevant information about the scene. Unlike conventional cameras, the event camera has advantages such as high temporal resolution, high dynamic range, low power consumption, and high bandwidth, which could allow improvements to be made to detection algorithms implemented at construction sites under hostile conditions [7].

This study considers the integration of the mobile robot ROSbot 2.0 developed by the company Husarion with the event camera (DAVIS 346) in the Gazebo emulation environment to detect workers, safety equipment, trees, tables and robots in the construction sector [8]. The mobile robot's displacement enables the acquisition of event information to apply detection algorithms based on YOLOv8 (You Look Only Once V8) and DETR (Detection Transformer).

II. RELATED WORK

Detection algorithms in robotic applications leveraging event cameras have facilitated the formulation of strategies for navigation, gesture recognition, and object manipulation [9], [10]. To this end, detection applications [11] present a manual event-driven human detection strategy. Additionally, databases of persons in different scenarios have been documented, including PAF [12] and pedestrian-SARI [13].

The study presented in [14] implements the surrogate gradient learning strategy of spike back-projection, parametric LIF, SpikingJelly frame and voxel-wise coding to train spiking neural networks (SNNs) with data from event cameras, improving the efficiency of object detection. Moreover, [15] proposes to fuse images and event voxel grids as input to a frame and event feature extractor network to overcome the detection efficiency presented by conventional cameras.

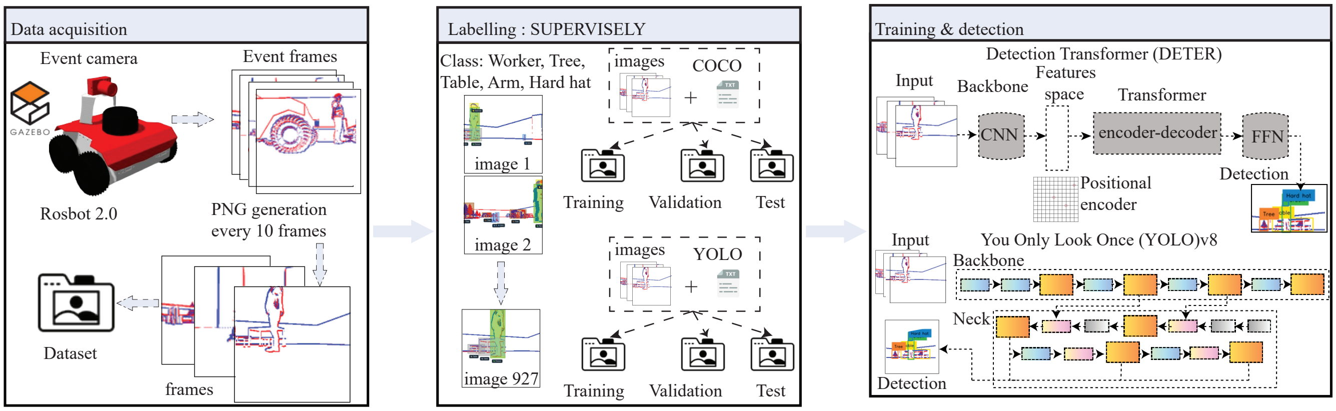


Fig. 1: The data acquisition process is performed through the Gazebo emulation software and manual labelling of five classes (i.e. tree, table, hard hat and worker) using Supervisely. Furthermore, the implemented algorithms DETR and YOLOv8 are trained, validated and tested.

Deep learning algorithms, however, cannot directly use information event cameras which come in the form of asynchronous event sequences. For this reason, the information obtained from the camera must be encoded in a tensor-based representation. One strategy is to transform the event stream into event frames. In this context, [16] reported the use of You Only Look Once (YOLO), which includes a small and large network algorithm to detect persons.

Moreover, event-based detection methods propose to represent events through hyperhistograms, which significantly improves the execution time of detection algorithms such as YOLOv5, Deformable-DETR and RetinaNet [17]. On the other hand, object detection is crucial for tracking tasks since it provides relevant information about the spatial location of objects in the scene. Hence, [18] reports a spiking transformer network (STNet) in this context. The proposed network extracts information from the global spatial domain and temporal signals through a transformer module and a spiking neural network (SNN). The cross-domain data fusion has outperformed current methods in accuracy and tracking speed.

III. Materials & Methods

Figure 1 shows a general overview of the proposed approach. First, data is extracted from the Gazebo emulation environment, which consists of the ROSbot 2.0 mobile robot, the event camera, and the objects to be detected (i.e., trees, tables, people and hard hats). The information obtained is then converted from events to event frames. Furthermore, manual labelling of five classes of objects based on the Supervisely software has been performed, generating a COCO and YOLO format to train the DETER and YOLOv8 detection algorithms, respectively. Finally, the implemented DETER and YOLOv8 architectures based on event frames are presented.

1) Data acquisition process: During a monitoring process involving a motion worker in a hard hat, trees, tables, and robots distributed in the environment, we generated data capturing the emulated playground environment of the Universidad de O'Higgins. In this context, event data was generated from an event camera plug-in in the Gazebo simulation software in ROS, using a DAVIS346 camera mounted on the ROSbot 2.0 mobile robot.

2) Computing hardware: The training and testing of the YOLOv8 and DETER detection algorithms were conducted with the following hardware specifications: CPU Intel® Core i3-12100F @4,30 GHz, GPU NVIDIA® GeForce RTX 3070 and 32 GB of RAM.

3) Pre-processing the dataset: The event information is recorded in $[x, y, p, t]$ tuples, where $[x, y]$ is the spatial location of the 2D event, $[p]$ is its polarity, either positive or negative, and $[t]$ is the time instant at which the change in pixel brightness is generated. Note that the event frames are recorded and updated every five milliseconds with a resolution of 346x260. The generated dataset is 927 event frames in PNG format, of which 647 frames are divided for training, 186 for validation and 94 for testing.

4) Deep Learning Algorithms:

- YOLOv8 is an improved version of the YOLOv5 architecture, which features an anchor-free model with decoupled heads in detection applications. The new version improves detection accuracy by integrating a C2F module and combining the sigmoid function for object scores. In addition, YOLOv8 uses CIoU and DFL functions based on bounding boxes to reduce losses. Finally, YOLOv8 in detection tasks presents higher accuracy, high speed and low computational cost.
- DETER is an architecture based on Transformers and convolutional neural networks (CNN), which achieves results comparable to fast recurrent convolutional neural networks (Faster R-CNN) in object detection applications. In this context, it uses features of a CNN and processes them through Transformers, including their spatial location, with a bipartite matching stage that improves detection confidence. The main advantages of DETER are its auto-detection, prioritisation of relevant features, and low computational cost.

5) YOLOv8 and DETER model training: Training YOLOv8 and DETER involves fitting from a previously trained model. YOLOv8 implements the YOLOv8-s model (s is small), employing the COCO data set and further retrained during 25 and 150 epochs using the event camera-generated data set with batch size 16 and 260x346 resolution. Meanwhile, DETER was trained from the residual neural network model (*Facebook/resnet-50*), which used the COCO data set for its prior training and then trained back with the event camera-acquired data set for the 25 and 150 epochs with a resolution of 260x346 and a batch size of 4.

6) Metrics: The detection algorithms are evaluated based on standard metrics used in supervised learning; these metrics include precision, recall, f1-score and accuracy; These are calculated according to equations (1), (2) and (3):

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{1}$$

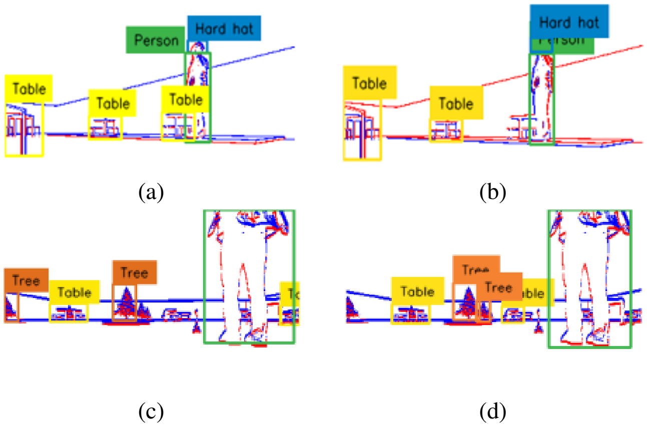
$$F_1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

where TP is true positive, TN represent true negative, FP defines false positive, and FN is false negative.

IV. Results

The experimental component consists of an emulated construction environment. This emulated environment was built from real data captured in the yard and laboratory of Robotics and Intelligent Systems (RISLAB) of the Universidad de O'Higgins, located in the commune of Rancagua, Chile. This environment has a worker (worker) with its respective helmet, labelled person and helmet, respectively. In addition, the scenario has 12 tables, six trees and two robotic arms. The scenario emulation consists of the movement of the mobile robot ROSbot 2.0 at a speed of 0.3 m/s, equipped with a DAVIS346 event camera. This emulation aims to detect details present in the environment.



(a) (b)

(c) (d)

Fig. 2: Object detection implemented deep learning algorithms from the event camera-based dataset; (a-c) ev-YOLOv8 detection in the event frames; (b-d) ev-DETER detection in the event frames.
Note: The results have been evaluated in event frames of five milliseconds, and the emulated scenario corresponds to the yard and Robotics and Intelligent Systems Laboratory at the Universidad de O'Higgins.

TABLE I: Analysis of performance metrics, including accuracy, recall, F1 score and precision, to evaluate the ev-YOLOv8 and ev-DETER detection algorithms in the test database.

Epochs	Model	Metrics			
		Precision	Recall	F1-Score	Accuracy
25	ev-YOLOv8	**0.892**	0.912	**0.902**	0.830
	ev-DETER	0.880	**0.922**	0.900	**0.846**
150	ev-YOLOv8	0.879	0.914	0.896	0.824
	ev-DETER	**0.921**	**0.934**	**0.927**	**0.884**

1) Qualitative results: We evaluate two methods, YOLOv8 and DETER, trained for handling data from event-based cameras. In the following, we will refer to them as ev-YOLOv8 and ev-DETER, methods that have demonstrated a high detection rate of workers, hard hats, trees, tables and robotic arms, as shown in Figure 2. Furthermore, we have noticed that the ev-YOLOv8 algorithm can detect an object even though the object is partially occluded (the bounding box covers part of the object). On the other hand, the ev-DETER algorithm can detect objects only when the element to be detected is within the central area of the bounding box of the label corresponding to the worker, robot, hard hat, table or tree.

2) Quantitative results: In Table 1, the results of the ev-YOLOv8 algorithm trained for 25 epochs have shown better accuracy and F1-Score metrics performance, outperforming ev-DETER by 1.2% and 0.2%, respectively. However, ev-DETER performs better in the recall and precision metrics, outperforming ev-YOLOv8 by 1% and 1.16%, respectively.

The evaluation results for a training of 150 epochs highlight the performance of the ev-DETER detection algorithm, as shown in Table 1a. In contrast to ev-YOLOv8, ev-DETER presents an increase in precision and accuracy of 4.2% and 6%, respectively. In addition, recall and F1-Scores metrics report a difference of more than 2% and 3.1%, respectively.

Table II reports the training and detection time of the ev-YOLOv8 and ev-DETER algorithms. Results report that the ev-YOLOv8 algorithm takes one-third of the training time of ev-DETER, which is approximately one minute per epoch, which is computationally expensive yet feasible. In contrast, ev-DETER takes 5.46 milliseconds of detection time compared to ev-YOLOv8, which requires 4.7 milliseconds, reducing the number of event frames processed by one second.

TABLE II: Training and detection time of the ev-YOLOv8 and ev-DETER convolutional neural network models.

Model	ModeTraining time by epochs [s]	Detection time [s]
ev-YOLOv8	**17.599**	**0.00467**
ev-DETER	60.339	0.00546

V. Discussion

Our main objective was to explore the potential of using event cameras to detect workers, safety equipment and objects (trees, tables) present in an emulated construction scenario. The system has been evaluated using two neural network models, ev-YOLOv8 and ev-DETER. The results show that ev-DETER is an efficient tool for object detection in construction sites due to its high precision, recall, F1-Score and accuracy compared to ev-Yolov8.

Integrating event cameras in construction processes could increase the potential of convolutional neural networks in detection applications. Implementing ev-DETER in the construction sector contributes to real-time monitoring of the location of robots and workers, ensuring cooperation and task optimisation. In addition, the use of ev-DETER for state detection of the construction process could be used to manage the workspace between workers and the physical built environment.

In summary, the ev-DETER algorithm enables efficient analysis based on neuromorphic technology, guaranteeing cooperative work between robots and workers in dynamic environments. Hence, ev-DETER could have a relevant role in detecting workers, security equipment and objects in the construction industry.

VI. CONCLUSIONS

The construction sector is a challenging environment for implementing computer vision algorithms based on deep neural networks using conventional cameras. The main challenges of a construction site are the abrupt changes in illumination, the presence of dust and the dynamic evolution of construction. These factors influence the accurate and efficient detection of workers, safety equipment and moving objects. This study evaluated a dynamic construction site detection system through deep learning techniques such as ev-YOLOv8 and ev-DETER using event cameras, which overcome the challenges presented by conventional cameras. The results show that ev-DETER performs better in precision, F1-Score, Recall and accuracy. In addition, it has been determined that the algorithm's performance improves with increasing training epochs.

Future work includes real-world evaluation of the proposed methods. The following next step includes combining detection algorithms and event cameras to explore the potential of neuromorphic technology in automated construction tasks involving the study of construction management and worker safety. This could increase the quality of projects and ensure the safety of workers in an accurate and timely manner.

ACKNOWLEDGMENT

This work is supported by ANID FONDECYT Postdoctorado 3240116

References

[1] F. Alsakka, S. Assaf, I. El-Chami, and M. Al-Hussein, "Computer vision applications in offsite construction," *Automation in Construction*, vol. 154, p. 104980, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580523002406

[2] M. Zhang, R. Xu, H. Wu, J. Pan, and X. Luo, "Human–robot collaboration for on-site construction," *Automation in Construction*, vol. 150, p. 104812, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580523000729

[3] R. Guamán-Rivera, A. Martínez-Rocamora, R. García-Alvarado, C. Muñoz-Sanguinetti, L. F. González-Böhme, and F. Auat-Cheein, "Recent developments and challenges of 3d-printed construction: A review of research fronts," *Buildings*, vol. 12, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2075-5309/12/2/229

[4] J. Li, Q. Miao, Z. Zou, H. Gao, L. Zhang, Z. Li, and N. Wang, "A review of computer vision-based monitoring approaches for construction workers' work-related behaviors," *IEEE Access*, vol. 12, pp. 7134–7155, 2024.

[5] R. Guamán-Rivera, O. Menéndez, T. Arevalo-Ramirez, K. Aro, A. Prado, R. García-Alvarado, and F. Auat-Cheein, "Assessment of convolutional neural networks for asset detection in dynamic automation construction environments," in *2023 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, 2023, pp. 1–6.

[6] F. Secci and A. Ceccarelli, "On failures of rgb cameras and their effects in autonomous driving applications," in *2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE)*, 2020, pp. 13–24.

[7] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.

[8] J. Kaiser, J. C. Vasquez Tieck, C. Hubschneider, P. Wolf, M. Weber, M. Hoff, A. Friedrich, K. Wojtasik, A. Roennau, R. Kohlhaas, R. Dillmann, and J. M. Zöllner, "Towards a framework for end-to-end control of a simulated vehicle with spiking neural networks," in *2016 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, 2016, pp. 127–134.

[9] A. Zujevs, M. Pudzs, V. Osadcuks, A. Ardavs, M. Galauskis, and J. Grundspenkis, "An event-based vision dataset for visual navigation tasks in agricultural environments," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 769–13 775.

[10] R. Verschae and I. Bugueno-Cordova, "Event-based gesture and facial expression recognition: A comparative analysis," *IEEE Access*, vol. 11, pp. 121 269–121 283, 2023.

[11] C. Boretti, P. Bich, F. Pareschi, L. Prono, R. Rovatti, and G. Setti, "Pedro: an event-based dataset for person detection in robotics," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 4065–4070.

[12] W. Shariff, M. S. Dilmaghani, P. Kielty, J. Lemley, M. A. Farooq, F. Khan, and P. Corcoran, "Neuromorphic driver monitoring systems: A computationally efficient proof-of-concept for driver distraction detection," *IEEE Open Journal of Vehicular Technology*, vol. 4, pp. 836–848, 2023.

[13] J. Wan, M. Xia, Z. Huang, L. Tian, X. Zheng, V. Chang, Y. Zhu, and H. Wang, "Event-based pedestrian detection using dynamic vision sensors," *Electronics*, vol. 10, no. 8, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/8/888

[14] L. Cordone, B. Miramond, and P. Thierion, "Object detection with spiking neural networks on automotive event data," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

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

[16] D. Przewlocka-Rus and T. Kryjak, "Power-of- two quantized yolo network for pedestrian detection with dynamic vision sensor," in *2023 26th Euromicro Conference on Digital System Design (DSD)*, 2023, pp. 39–45.

[17] Y. Peng, Y. Zhang, P. Xiao, X. Sun, and F. Wu, "Better and faster: Adaptive event conversion for event-based object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 2056–2064, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25298

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