ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0367
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

DTR: A Unified Detection-Tracking-Re-identification Framework for Dynamic Worker Monitoring in Construction Sites

Nasrullah Khan[1], Syed Farhan Alam Zaidi[2], Aqsa Sabir[3], Muhammad Sibtain Abbas[4], Rahat Hussain[5], Chansik Park[6], Dongmin Lee[7]*

[1] *Department of Architectural Engineering, Chung-Ang University, Seoul, South Korea,* E-mail address: nasazzam@cau.ac.kr
[2] *Department of Computer Science, Chung-Ang University, Seoul, South Korea,* E-mail address: syedfarhanalam1993@gmail.com
[3] *Department of Computer Science, Chung-Ang University, Seoul, South Korea,* E-mail address: aqsasabir786@gmail.com
[4] *Department of Architectural Engineering, Chung-Ang University, Seoul, South Korea,* E-mail address: muhammadsibtain@cau.ac.kr
[5] *Department of Architectural Engineering, Chung-Ang University, Seoul, South Korea,* E-mail address: rahat4hussain@gmail.com
[6]*Professor, Department of Architecture and Building Science, Chung-Ang University*, *E-mail address: cpark@cau.ac.kr*
[7] *Assistant Professor, Department of Architecture and Building Science, Chung-Ang University, E-mail address: dmlee@cau.ac.kr*

Abstract:
The detection and tracking of construction workers in building sites generate valuable data on unsafe behavior, work productivity, and construction progress. Many computer vision-based tracking approaches have been investigated and their capabilities for tracking construction workers have been tested. However, the dynamic nature of real-world construction environments, where workers wear similar outfits and move around in often cluttered and occluded regions, has severely limited the accuracy of these methods. Herein, to enhance the performance of vision-based tracking, a new framework is proposed which seamlessly integrates three computer vision components: detection, tracking, and re-identification (DTR). In DTR, a tracking algorithm continuously tracks identified workers using a detector and tracker in combination. Then, a re-identification model extracts visual features and utilizes them as appearance descriptors in subsequent frames during tracking. Empirical results demonstrate that the proposed method has excellent multi-object-tracking accuracy with better accuracy than an existing approach. The DTR framework can efficiently and accurately monitor workers, ensuring safer and more productive dynamic work environments.

Keywords: Worker tracking, Computer vision, Re-identification (ReID), Deep Learning, Dynamic worker monitoring.

1. INTRODUCTION

Workers can be tracked for productivity analysis and safety monitoring construction sites [1]. For example, knowing the exact location of workers enables resource distribution optimization and efficient evacuations during emergency situations. Camera-based systems are widely available and have substantial benefits due to the information-rich visual data they can collect and process and their ease of deployment in most construction scenarios[2-4]. With recent advancements in technology, surveillance

systems can incorporate artificial intelligence (AI) detectors based on deep learning [5] to localize workers in different construction environments for specific use cases [2, 5-7]. However, although detectors can identify workers, they do not link them across successive frames and are susceptible to false detections, limiting practical applications e.g. productivity analysis. In some cases, the detectors are combined with motion-based tracking algorithms [8, 9] to locate the same object in the current frame and track its position in the succeeding frames. Moreover, appearance features can be added as an additional input or the detection can be based on confidence scores. Combining detection and tracking addresses the limitations of detectors in accounting for workers across multiple frames and reducing false detections, thus improving the performance of vision-based methods in applications that require continuous monitoring of workers over time.

In a dynamic construction site, worker tracking encounters unique challenges due to the inherently unpredictable environment. Unlike static scenarios, construction sites are characterized by fluctuating lighting conditions, worker uniformity in terms of clothing and safety gear, and frequent occlusions caused by machinery and materials. Constant worker movement and diverse task execution add further complexity. These dynamic elements significantly impact traditional detection and tracking systems, leading to misdetections arising from poor visibility and mismatched associations due to similar appearances. Consequently, tracked trajectories are lost and become unreliable, hindering the effectiveness of the system. To address these issues, in this study, an integrated detection-tracking-re-identification (DTR) framework is proposed. This comprehensive strategy seamlessly integrates three key elements: a detector, a tracking algorithm, and a re-identification (ReID) model for re-identifying tracked objects across different frames and to recover lost tracks [10]. The components of the integrated system work synergistically: (a) the detector locates workers, (b) the tracker associates detected workers across frames, and (c) the ReID model assists the tracker by recovering identities after temporary occlusions or other factors.

This paper is organized as follows. In Section 2, the current state of research on object detection and tracking as applied to the construction industry is reviewed. In Section 3 the technical details for the proposed computer vision-based detection and tracking system are presented. Section 4 analyzes the tracking performance of the proposed DTR framework, and Section 5 summarizes the main conclusions from the study as well as provides the scope of future research.

2. Literature review

Object detection and tracking are fundamental tasks in computer vision and have a wide range of applications, including video surveillance, autonomous vehicles, and robotics [5, 11-13]. Son and Kim [6] developed an integrated construction worker detection and tracking scheme for real-time workspace monitoring to enhance safety in construction sites. They used a siamese neural network but the system had difficulty in locating workers that were occluded by more than 50% in poorly illuminated scenarios leading to mismatches and swapped target identities. Teizer et al. [11] proposed a classification model involving a three-stage process to collectively locate, track, and re-identify workers while ensuring compliance with personal protective equipment (PPE) regulations. However, despite its potential benefits, the complexity of this system makes it unfeasible in practice. Specifically, the time required to process one pair of frames using their method is longer than the interval between frames, which could lead to significant delays in detecting and tracking workers. Similarly, Angah et al. [3] utilized a deep learning model for detection and gradient-based re-matching. However, the scale variation and visual similarity between workers wearing PPE make it challenging to re-identify workers only using gradient-based

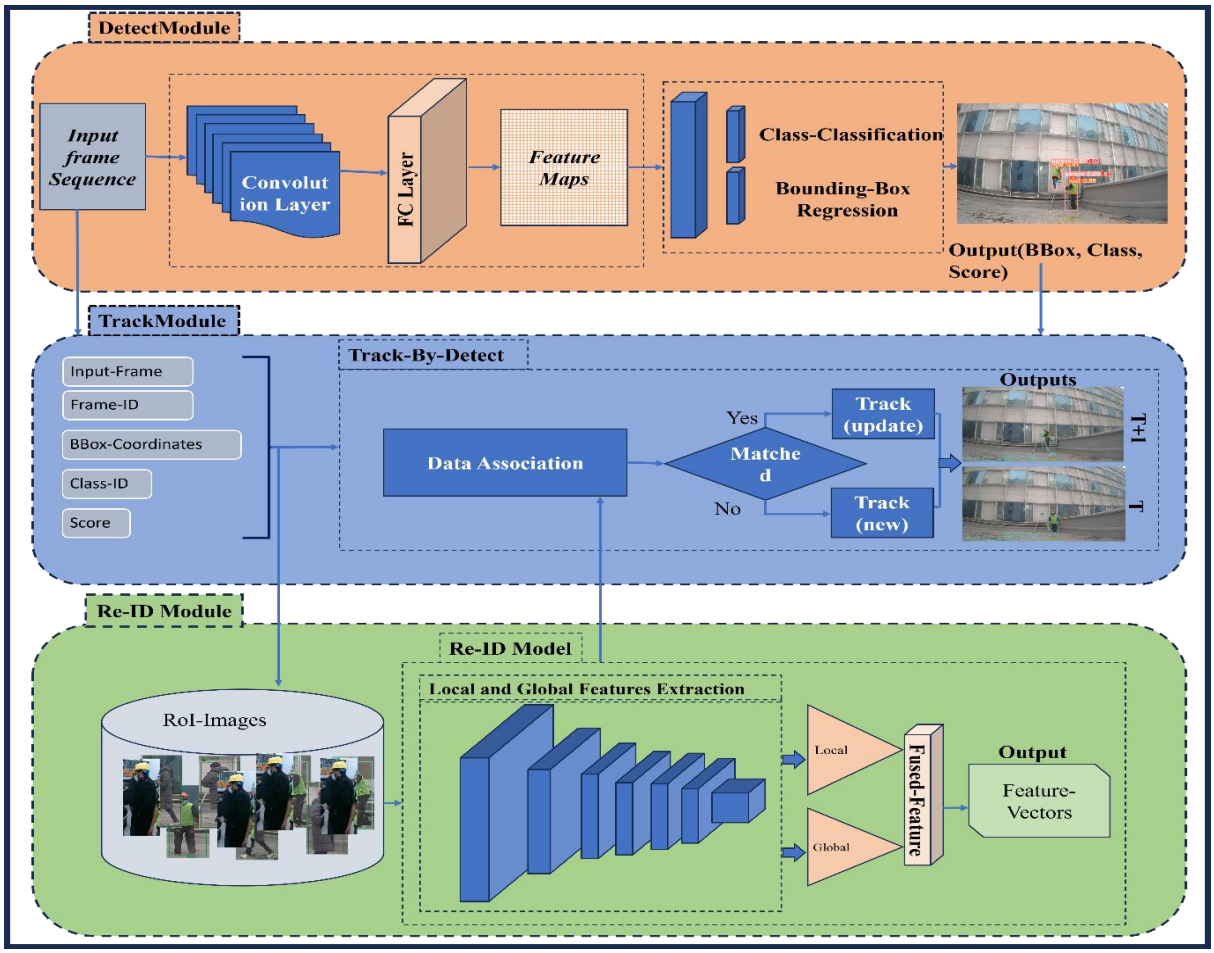
matching. Also, dynamic backgrounds have high pixel variability that is inconsistent and construction scenes lack collective pixel movement. An important performance metric is the intersection over union (IoU) score and the IoU may be too low in these scenarios and deteriorate the re-matching process for the target worker. Teizer et al. [4] investigated tracking algorithms such as density mean-shift and Bayesian segmentation, highlighting challenges like occlusion and the need for vast training datasets. Neuhuasen et al. [14] explored bird's-eye view cameras, proposing adapted pedestrian detection algorithms but the existence of partial occlusions necessitated robust pose estimation. Yang et al. [15] proposed a machine learning-based approach with a parameterized feature bank to handle multiple workers and appearance variations. However, this strategy suffered from poor initialization with identification sample templates. Existing solutions address isolated challenges based on some tracking methods, neglecting the interplay of dynamic factors like lighting, occlusion, and worker similarity. Additionally, identity mismatches from occlusions and similarities remain unaddressed, and solutions lack generalizability across diverse construction scenarios. Consequently, valuable information beyond target location are missing, impacting resource allocation, safety, and project management applications.

Further research on an integrated solution that accounts for all dynamic factors, effectively addresses persistent target mismatches, and ensures broad applicability is required. This will not only improve tracking accuracy but also unlock the true potential of worker tracking data to transform the construction industry. Despite previous efforts in worker tracking, mistaken identity (ID) errors due to target switching are prevalent, leading to lost tracks and compromised data. In practical environments, mistaken IDs are often due to occlusion and cause mismatches and data gaps, jeopardizing worker safety, operational efficiency, and the utility of advanced tracking technologies. Lost tracks hinder reliable data analysis for task optimization and resource allocation. Also, lost tracks and switching mask potential safety risks, especially in hazardous environments. This study aims to develop a computer vision-based tracking system that minimizes identity switching and lost tracks, paving the way for their successful implementation in dynamic workplaces.

3. Proposed Framework

To eliminate ID switching due to missed detections or erroneous matching when tracking workers in construction sites, a variety of strategies have been proposed, some emphasizing complexity of architecture to improve performance and others emphasizing execution time of the system. The former focused on visual traits, whereas the latter employed IoU-based matching. In this study, a dynamic worker tracking system that combines visual-features with IoU-based matching was developed and validated. The proposed DTR framework integrates three modules: DetectModule, TrackModule, and Re-IDModule as shown in Fig 1.

Figure 1 : DTR: Detection-Tracking-Re-identification framework for construction worker monitoring



DTR integrates detection and tracking techniques assisted by re-identification model features to maintain the identity consistency of objects across sequential frames. The process begins with an *Input Frame Sequence*, where a stream of frames is fed into the *DetectModule* consisting of an object detector. The frames undergo an initial transformation in the Convolution Layer, which applies convolutional operations to extract spatial hierarchies of features. The output from the fully connected (FC) layer is a set of Feature Maps, and these features are passed to the Neck and Head of the object detector model so that different objects can be detected and distinguished. The output of the detection model consists of bounding-box coordinates (BBox), confidence score (score), and Class ID. In the second step, TrackModule takes the output stream from DetectModule and adds track ID to the detection input which helps in tracking the object of interest in consecutive frames. The data association is particularly challenging for a construction site due to the frequent occlusion and interaction between workers and machinery. Our proposed framework is categorized as a track-by-detect technique followed by Re-ID-based tracking, where the BBox is selected by score and assisted by Re-IDModule feature vectors to associate detected objects across consecutive frames. The subset *Re-IDModule* focuses on RoI-Images (Regions of Interest), where objects within the frame are isolated or cropped with the help of BBox coordinates and the original frame of the detected object. These cropped regions are passed to a Re-Identication model which extracts Local and Global Features. Both Local and Global Features are fused and converted to a feature space which assists in matching tracks using a feature vector similarity matrix. The cosine calculations compare newly extracted feature vectors with existing ones, facilitating the decision-making process for worker re-identification. Local features pertain to the individual aspects of a worker, such as a unique color or shape within a bounded region, while global features refer to the holistic attributes of an individual worker. The integration of both local and global features ensures robustness in feature vectors, enabling the system to accurately track objects despite variations in scale, orientation, and illumination. The DTR framework leverages the features to assist which objects are consistent throughout the

sequence, marking them as 'matched' even after occlusion or other factors affecting track trajectories. with any discrepancies resulting in new objects being initiated into tracking. The output is twofold: Track (update), where existing objects such as workers are continuously tracked across frames updated with new positional data, and Track (new), which involves newly detected workers that have just entered the field of view and require the initiation of a new track. DTR overcomes the unique challenges of object tracking in the construction industry and can be tailored for different construction environments.

4. Case study

4.1 Case scenario

We validated the proposed framework by annotating testing videos first used by Konstantinou et al. [16] involving different case scenarios. The videos were annotated in MOT format [17], a standard for evaluating tracking performance. The test scenarios encompassed diverse challenges like "background clutter is rampant", "with PPE equipment", and "materials easily mistaken for workers". Occlusion is frequent, with workers obscured by machinery, walls, or each other. Scale variation comes into play due to varying distances from the cameras of workers. Posture variation is amplified by diverse activities like bending, climbing, or carrying objects. Abrupt movement is commonplace, with workers quickly maneuvering around the site or operating heavy machinery. The dynamic and often chaotic nature of construction environments create a complex puzzle for algorithms and demand specialized approaches. Such scenarios reflect real-world applications where the accuracy of tracking depends on overcoming these hurdles. Table 1 summarizes the characteristics of the annotated videos. Representative results are shown in Fig. 2.

Table 1: Annotated videos used to evaluate the proposed dynamic tracking framework.

Video #	Frames	Worker	Challenges for Tracking
Vid1	227	1	Background Clutter, Abrupt Movement
Vid2	258	3	Background Clutter, Occlusion
Vid3	357	4	Occlusion, Scale Variation
Vid4	258	6	Posture Variation, Scale Variation

4.2 Experimental Setup

The authors performed tracking experiments using a Windows OS and equipped with a GeForce Nvidia RTX 3090 Ti 24GB VRAM GPU and 32GB of RAM. The proposed system was implemented using Pytorch, a deep-learning framework for Python.

4.3 Evaluation Metric

The performance of the proposed dynamic tracking framework was quantified using the multi-object-tracking accuracy (MOTA) metric. MOTA takes into account misses, false positives, and identity switching in the tracked objects using the following equation:

$$MOTA = 1 - \frac{\text{FN}+\text{FP}+ID_{sw}}{\text{GT}} \quad (1)$$

where FN represents the number of false negatives (missed detections), FP represents the number of false positives (incorrectly identified detections), IDsw represents the number of identity switches, and GT represents the total number of ground truth objects. It is the sum of all errors made by the tracker over all frames, averaged by the total number of ground-truth objects. MOTA is similar to the accuracy metrics widely used in other domains and gives a very intuitive measure of the tracking performance. In multi-object tracking, each unique object is assigned an ID that should remain constant across video frames. However, ID switching can occur when the tracker incorrectly associates the ID of one object to a different object or a new ID is assigned to the object in the sequence of frames. This violates the uniqueness of the ID assignment and ID switching negatively impacts the MOTA score. By tying the identity more closely to the real-world appearance of the object, rather than an internal label, we improved the tracking stability and reduced errors like ID switching. This leads to a more accurate frame-to-frame tracking of multiple objects, and consequently, a higher MOTA score.

4.4 Experimental Results

The proposed framework utilizes the YOLOX [18] detector and ByteTrack [19] tracking algorithm in conjunction with the OSNet Re-ID [20] model as a feature assister. This combination enables the framework to achieve superior tracking performance. Table 2 presents the empirical results obtained using the smallest and largest variants of YOLOX using the experimental setup described in Section 4.1. The detection models were trained on a publicly available worker dataset [21], resulting in a mean average precision (mAP) of 92.1 for YOLOX-NANO and 95.3 for YOLOX-L. A pretrained ReID model [20] was used to extract worker features. The results demonstrate a clear correlation between mAP and visual feature assistance. Higher mAP values lead higher MOTAs and a reduction in the number of IDs, with an inverse effect on the average FPS (Frames-Per-Second). The proposed framework, which incorporates the ReID model, outperforms the existing tracking system (ByteTrack) in terms of tracking performance (Table 2). Figure 2 shows inference results using a drone with our proposed ReID and without ReID. In Fig 2 (a) the tracking starts without ReID, and after 90% occlusion by the worker on the scaffolding the ID is not recovered in Fig 2 (b) (light purple BBox). On the other hand, with ReID, the ID is recovered in the next frame after occlusion (Fig (c)-(d)). To optimize inferencing we used ONNX runtime quantization [22] as shown in Fig 2 (e) to achieve ~30 fps on YOLOX-NANO with Person ReID.

Table 2 : MOTA results using different YOLOX versions with only the tracking algorithm and with ReID feature assist.

Model	**MOTA**
YOLOX-NANO	74.6%
YOLOX-NANO (Re-ID)	77.8%

YOLOX-L	78.9%
YOLOX-L (Re-ID)	81.3%

.

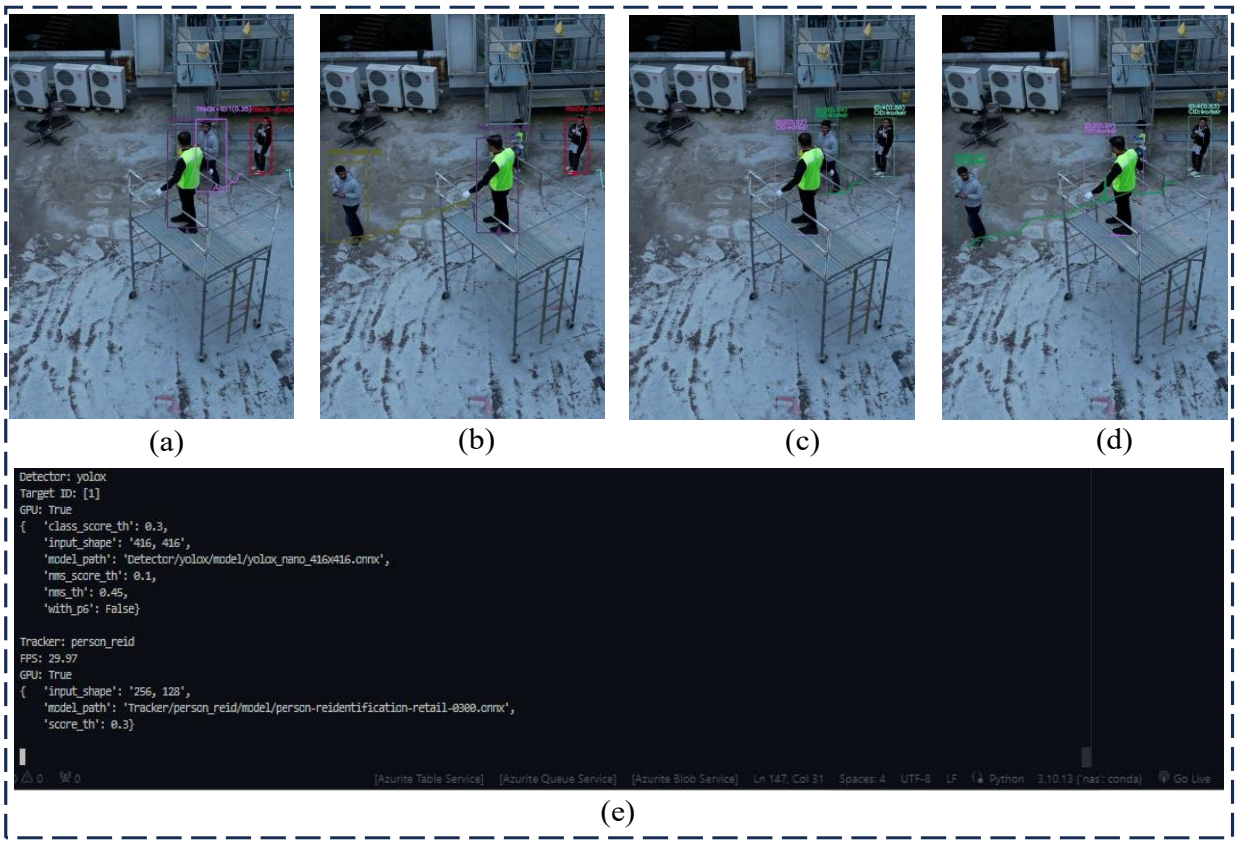


Figure 2 : Image tracking results (a)(b) without incorporating the proposed approach using the object detector ByteTrack, and (c)(d) incorporating the proposed approach with ReID. From the images, it is clear that the proposed approach is highly effective in preserving the identity of the person even when there is 90% occlusion. In contrast, without ReID there is a significant change in the identity of the person as represented by the shift in color (b). (e) ONNX runtime quantization to optimize inferencing.

5. Conclusion and Future Work

In this study, the authors aimed to efficiently detect and track workers in construction sites with challenging conditions such as occlusion and variations in illumination which cause ID switching during tracking. To achieve this, an integrated detect, track, and re-identification (DTR) framework was developed. This framework implemented a track-by-detect approach in tracking the worker, assisted by ReID model features of the detected worker when tracking. The framework is designed to be flexible and adaptable to different job sites depending on the use case. Post-processing techniques can then be implemented to analyze the tracking data and identify specific visual cues. These visual cues can be used for practical applications such as identifying potential PPE violations in dangerous areas such as workers not wearing hardhats or workers without high-visibility vests. These unsafe practices can then be flagged automatically so that the relevant safety intervention can be made. In the future authors will validate the framework using larger custom tracking datasets from the construction domain to generate a benchmark tracking dataset that can help validate the performance of tracking systems in the construction industry.

ACKNOWLEDGMENTS

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00217322).

REFERENCES

1. Teizer, J., T. Cheng, and Y. Fang, *Location tracking and data visualization technology to advance construction ironworkers' education and training in safety and productivity.* Automation in construction, 2013. **35**: p. 53-68.
2. Zhang, M., R. Shi, and Z. Yang, *A critical review of vision-based occupational health and safety monitoring of construction site workers.* Safety science, 2020. **126**: p. 104658.
3. Angah, O. and A.Y. Chen, *Tracking multiple construction workers through deep learning and the gradient based method with re-matching based on multi-object tracking accuracy.* Automation in Construction, 2020. **119**: p. 103308.
4. Teizer, J. and P.A. Vela, *Personnel tracking on construction sites using video cameras.* Advanced Engineering Informatics, 2009. **23**(4): p. 452-462.
5. Xu, S., et al., *Computer vision techniques in construction: a critical review.* Archives of Computational Methods in Engineering, 2021. **28**: p. 3383-3397.
6. Xu, Z., J. Huang, and K. Huang, *A novel computer vision-based approach for monitoring safety harness use in construction.* IET Image Processing, 2023. **17**(4): p. 1071-1085.
7. Zhong, B., et al., *Mapping computer vision research in construction: Developments, knowledge gaps and implications for research.* Automation in Construction, 2019. **107**: p. 102919.
8. Son, H. and C. Kim, *Integrated worker detection and tracking for the safe operation of construction machinery.* Automation in Construction, 2021. **126**: p. 103670.
9. Park, M.-W. and I. Brilakis, *Continuous localization of construction workers via integration of detection and tracking.* Automation in Construction, 2016. **72**: p. 129-142.
10. Brown, L., et al., *Performance evaluation of surveillance systems under varying conditions.* 2023.
11. Dai, Y., et al., *A survey of detection-based video multi-object tracking.* Displays, 2022. **75**: p. 102317.
12. Amin, S.U., et al., *Deep learning based active learning technique for data annotation and improve the overall performance of classification models.* Expert Systems with Applications, 2023. **228**: p. 120391.
13. Amin, S.U., et al., *An automated chest X-ray analysis for COVID-19, tuberculosis, and pneumonia employing ensemble learning approach.* Biomedical Signal Processing and Control, 2024. **87**: p. 105408.
14. Neuhausen, M., J. Teizer, and M. König. *Construction worker detection and tracking in bird's-eye view camera images*. in *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*. 2018. IAARC Publications.
15. Yang, J., et al., *Tracking multiple workers on construction sites using video cameras.* Advanced Engineering Informatics, 2010. **24**(4): p. 428-434.
16. Konstantinou, E., J. Lasenby, and I. Brilakis, *Adaptive computer vision-based 2D tracking of workers in complex environments.* Automation in Construction, 2019. **103**: p. 168-184.
17. Milan, A., et al., *MOT16: A benchmark for multi-object tracking.* arXiv preprint arXiv:1603.00831, 2016.
18. Ge, Z., et al., *Yolox: Exceeding yolo series in 2021*. arXiv preprint arXiv:2107.08430, 2021.
19. Zhang, Y., et al. *Bytetrack: Multi-object tracking by associating every detection box*. in *European Conference on Computer Vision*. 2022. Springer.
20. Zhou, K., et al. *Omni-scale feature learning for person re-identification*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
21. Hard-Hat. *PPE-RD-v1 Dataset*. Roboflow Universe [Open Source Dataset] 2023 Available from: https://universe.roboflow.com/hard-hat-ihqrk/ppe-rd-v1.
22. developers, O.R. *ONNX Runtime*. 2021; Available from: https://onnxruntime.ai/.