

Article

Optimizing Helmet Use Detection in Construction Sites via Fuzzy Logic-Based State Tracking

Xiaoxiong Zhou ¹, Xuejun Jia ^{1,2}, Jian Bai ^{1,2}, Xiang Lv ², Xiaodong Lv ^{1,3} and Guangming Zhang ^{1,*} 

¹ College of Electrical Engineering and Control Science, Nanjing Tech University, Nanjing 211816, China; zhouxx@njtech.edu.cn (X.Z.); jxj1147208@sina.com (X.J.); baijian@njtech.edu.cn (J.B.); lxd@ycit.edu.cn (X.L.)

² China Construction Second Engineering Bureau Co., Ltd., Beijing 100160, China; lvxiang@mails.jlu.edu.cn

³ School of Electrical Engineering, Yancheng Institute of Technology, Yancheng 224051, China

* Correspondence: zgm@njtech.edu.cn

Abstract

Automated safety monitoring on construction sites requires precise helmet-status detection and robust multi-object tracking in long, occlusion-rich video sequences. This study proposes a two-stage framework: (i) a YOLOv5 model enhanced with self-adaptive coordinate attention (SACA), which incorporates coordinate-aware contextual information and reweights spatial-channel responses to emphasize head-region cues—SACA modules are integrated into the backbone to improve small-object discrimination while maintaining computational efficiency; and (ii) a DeepSORT tracker equipped with fuzzy-logic gating and temporally consistent update rules that fuse short-term historical information to stabilize trajectories and suppress identity fragmentation. On challenging real-world video footage, the proposed detector achieved a mAP@0.5 of 0.940, surpassing YOLOv8 (0.919) and YOLOv9 (0.924). The tracker attained a MOTA of 90.5% and an IDF1 of 84.2%, with only five identity switches, outperforming YOLOv8 + StrongSORT (85.2%, 80.3%, 12) and YOLOv9 + BoT-SORT (88.1%, 83.0%, 10). Ablation experiments attribute the detection gains primarily to SACA and demonstrate that the temporal consistency rules effectively bridge short-term dropouts, reducing missed detections and identity fragmentation under severe occlusion, varied illumination, and camera motion. The proposed system thus provides accurate, low-switch helmet monitoring suitable for real-time deployment in complex construction environments.

Keywords: helmet detection; YOLOv5; DeepSORT; fuzzy logic



Academic Editor: Nikolay L. Kazanskiy

Received: 6 September 2025

Revised: 13 October 2025

Accepted: 17 October 2025

Published: 21 October 2025

Citation: Zhou, X.; Jia, X.; Bai, J.; Lv, X.; Lv, X.; Zhang, G. Optimizing Helmet Use Detection in Construction Sites via Fuzzy Logic-Based State Tracking. *Sensors* **2025**, *25*, 6487. <https://doi.org/10.3390/s25206487>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Construction remains one of the most hazardous industries, with falls and head injuries a leading cause of severe and fatal incidents [1,2]. Proactive, site-wide monitoring of workers' helmet-wearing status is therefore essential. In practice, this requires coupling per-frame helmet-status detection with multi-object tracking (MOT) in long, cluttered videos captured by fixed site cameras, under real-time constraints.

Compared with generic pedestrian scenes, construction sites pose distinctive vision challenges (Figure 1): (i) an uneven distribution of workers that are frequently occluded by tools, scaffolds, or body pose, making the head-helmet boundary hard to resolve [3,4]; (ii) visual confounds—caps, hoods, and high-visibility vests with similar colors—produce hard negatives [5]; (iii) motion blur and rapid illumination changes from machinery and outdoor conditions degrade appearance features [6,7]; and (iv) long-duration identity

persistence is required as crews enter/exit and re-appear after occlusion, which commonly fragments tracks and triggers identity switches [8,9].

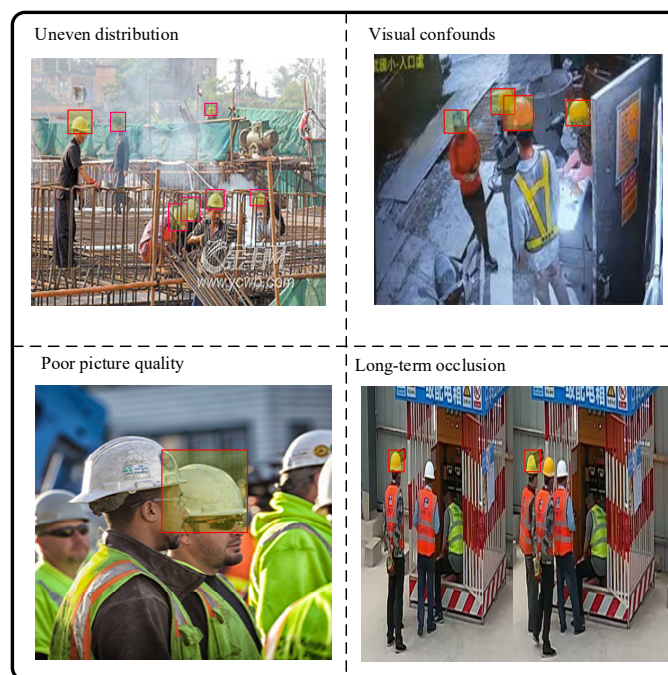


Figure 1. Challenges of helmet wearing detection on construction sites.

Although many countries have enacted laws requiring construction workers to wear helmets, traditional manual supervision methods have numerous limitations. First, manual inspections often fail to cover the entire construction site and require a large workforce for real-time monitoring, which is both costly and inefficient [10]. Additionally, traditional video surveillance systems typically rely on simple image processing methods that work well in static environments but struggle in complex dynamic environments. In particular, when workers move quickly, change their posture, or experience partial occlusion, these methods often suffer from reduced accuracy, leading to false positives or missed detections [11].

With the rapid development of artificial intelligence and computer vision technologies, video-based safety helmet detection has gradually become an important tool for safety management on construction sites. Most existing helmet detection methods rely on traditional image processing and deep learning technologies such as the YOLO series of object detection models. YOLOv5, as an efficient real-time object detection framework, has an excellent performance in many computer vision tasks. However, in the complex and dynamic environment of construction sites, the traditional YOLOv5 model still faces limitations when dealing with fast-moving workers, occlusion, and complex backgrounds [12]. Furthermore, object tracking algorithms such as DeepSORT are widely used for multi-object tracking tasks but still encounter issues with recognition errors and identity switches when dealing with highly similar targets or occlusions [13].

To address these challenges, this study proposes an improved helmet detection and tracking framework based on YOLOv5 and a fuzzy logic-enhanced DeepSORT algorithm. The main innovations of this work are reflected in the following two key aspects:

(1) Improved YOLOv5 for helmet detection: To enhance detection accuracy in dynamic construction environments, an adaptive spatial and channel attention mechanism was integrated into YOLOv5. This enabled the model to dynamically adjust its attention regions based on contextual information, thereby improving the precise identification of

targets. Particularly in cases where the workers' heads are occluded or moving rapidly, the improved YOLOv5 demonstrated superior capability in detecting helmet presence.

(2) In the tracking phase, DeepSORT was enhanced with a fuzzy logic module that dynamically adjusted the target-matching weights by combining intersection over union (IoU) and appearance features. When targets experience occlusion or visual similarity, fuzzy logic enables a more flexible and adaptive matching strategy, improving the tracking accuracy and stability. Furthermore, to address occlusion and target loss in dynamic environments, a temporal processing-based multi-frame information fusion method was introduced. By integrating information from consecutive frames, this approach mitigates the impact of false or missed detections in individual frames, thereby enhancing the model's robustness and continuity in complex scenarios.

The remainder of this paper is organized as follows. Section 2 reviews related research progress; Section 3 presents the proposed method and technical framework, with a focus on the application of fuzzy logic to the DeepSORT tracking algorithm; Section 4 describes the dataset and experimental design; Section 5 reports the experimental results; and Section 6 discusses the findings, summarizes the contributions, and outlines future research directions.

2. Related Works

2.1. Helmet Wearing Detection

Compared with generic pedestrian detection, helmet-wearing detection on active construction sites is harder for three reasons: the effective head region is small and frequently truncated or occluded by tools and scaffolds; caps, hoods, and high-visibility garments introduce look-alike distractors; and illumination shifts and motion blur are common in long, cluttered videos [14,15]. Early pipelines built on color/shape heuristics, background subtraction, or classical descriptors (e.g., HOG/SVM) are lightweight but degrade sharply under pose change, scaffold clutter, and lighting variation, producing unstable precision/recall in real sites [16–18]. Modern CNN detectors—both two-stage and one-stage, with YOLO-family models widely used for real-time PPE analytics—improve the per-frame accuracy through multiscale feature pyramids, dense heads, and better optimization [19–22]; Attention and transformer augmentations further emphasize the small head region and suppress background noise [23–26]. Progress has also come from anchor-free heads, improved label-assignment strategies, and focal/IoU-style losses that raise the sensitivity to small objects and address class imbalance [27–29] as well as feature alignment/refinement modules and hard-negative mining that reduce misalignment and PPE color confounds [30–32]. Nevertheless, per-frame decisions remain brittle under partial visibility, crowded scenes, and rapid lighting/pose changes, which continue to trigger false alarms and misses in practice [33–35].

Beyond detecting helmets *per se*, many studies infer helmet-wearing status by encoding the head–helmet spatial relationship—via pose/keypoint-guided head regions, part-based heads, or graph/constraint formulations—so that status depends on the relative geometry rather than the incidental overlap of person and helmet boxes [36–38]. Segmentation-based or hybrid detect-then-segment designs offer finer boundary cues but can increase the latency and annotation cost [39]. Robustness across projects and cameras remains challenging due to domain shift in garments, backgrounds, and viewpoints; to improve generalization, works have explored extensive augmentation (color/blur/weather/copy-paste), synthetic PPE generation, and adaptation/regularization (self-training, style/feature alignment) [40–42]. Even so, evaluations often underrepresent minute-long, occlusion-rich videos typical of active sites, where severe truncation and repeated re-appearances accumulate detection errors [43].

2.2. Object Tracking Algorithms

Compared with generic pedestrian tracking, monitoring helmet wearing on active construction sites poses domain-specific difficulties: workers undergo long, heavy occlusions, operate in crowded scenes with visually similar PPE, and appear in minute-long videos with re-entries and lighting changes. Most systems adopt a tracking-by-detection paradigm in which per-frame detections are linked across time [44]. DeepSORT-style trackers couple a linear-Gaussian motion model with Hungarian one-to-one assignment and a global appearance embedding [45]. These design choices explain both strengths and failure modes: the constant-velocity prior stabilizes short gaps but is violated by abrupt human maneuvers, while the Hungarian solver commits globally based on a noisy cost matrix, so when detections are partially overlapped or missing, near-ties in IoU/appearance can swap identities at crossings. ByteTrack-like strategies that anchor association on high-confidence detections reduce sensitivity to low-quality boxes, but thresholding suppresses genuine but low-score observations during blur/occlusion, creating track fragmentation that later promotes switches on re-appearance [46]. In addition, re-ID embeddings trained on everyday attire exhibit a domain gap in PPE-rich scenes: the head ROI is small, salient colors are shared across helmets/vests, and background clutter reduces inter-class separability, which lowers appearance discrimination after short misses [47]. Taken together, identity errors emerge not because these trackers are “insufficient”, but because their probabilistic and assignment assumptions (linear motion, stationary appearance, reliable detections) are routinely violated on construction footage.

To mitigate these violations, recent work advances four complementary directions. (i) Stronger association metrics. By fusing geometric overlap with appearance and adding occlusion-aware gating, the cost surface becomes better conditioned, reducing tie cases that trigger Hungarian misassignments under partial visibility [48]. (ii) Sequence-level modeling. Tracklet linking, short-window temporal smoothing, or graph/RNN aggregation explicitly propagate evidence across frames, repairing gaps created when detectors down-score blurred/occluded heads [49]. (iii) PPE-aware re-identification. Head/upper body-focused embeddings or part cues increase the between-identity margin in scenes where full-body clothing is homogeneous or safety gear is similar, improving post-occlusion re-association [50]. (iv) Real-time efficiency. Lightweight feature extractors and streamlined association preserve the above gains within edge latency budgets common to site deployments [51]. Parallel to these, fuzzy-logic controllers treat IoU and appearance similarity as uncertain linguistic variables and adaptively re-weight them when either cue is unreliable, thereby smoothing decisions near ambiguous boundaries [52]; auxiliary cues such as head/pose keypoints can further constrain association by injecting geometric consistency when boxes are truncated [53]. Despite progress, many evaluations still emphasize short clips with mild occlusion, so accumulated errors from long disappearances and re-entries remain under-measured; in practice, identity continuity is the dominant failure mode precisely because assumption violations persist over long horizons.

3. Methodology

This section details the core technical methods for helmet detection and worker tracking on construction sites, with a specific focus on the design and functionality of each module. Figure 2 shows the overall framework of the algorithm proposed in this paper, which mainly consists of two key parts: (i) helmet detection and worker tracking based on the improved YOLOv5, and (ii) the helmet-worker matching process.

First, the improved YOLOv5 model is used for helmet detection, with a particular emphasis on the SACA mechanism. This attention mechanism helps guide the model to focus more on the head region, enabling the accurate identification of helmet usage,

which significantly improves the robustness and accuracy of helmet detection. Next, the DeepSORT object tracking algorithm, combined with fuzzy logic, was used for worker matching and continuous tracking. Especially in situations with highly similar targets or partial occlusion, this method demonstrates high flexibility and stability. Finally, to further enhance the model's adaptability in dynamic environments, a temporal processing mechanism for multi-frame information integration was introduced. This smooths the tracking results, ensuring high accuracy and consistency throughout the detection and tracking process.

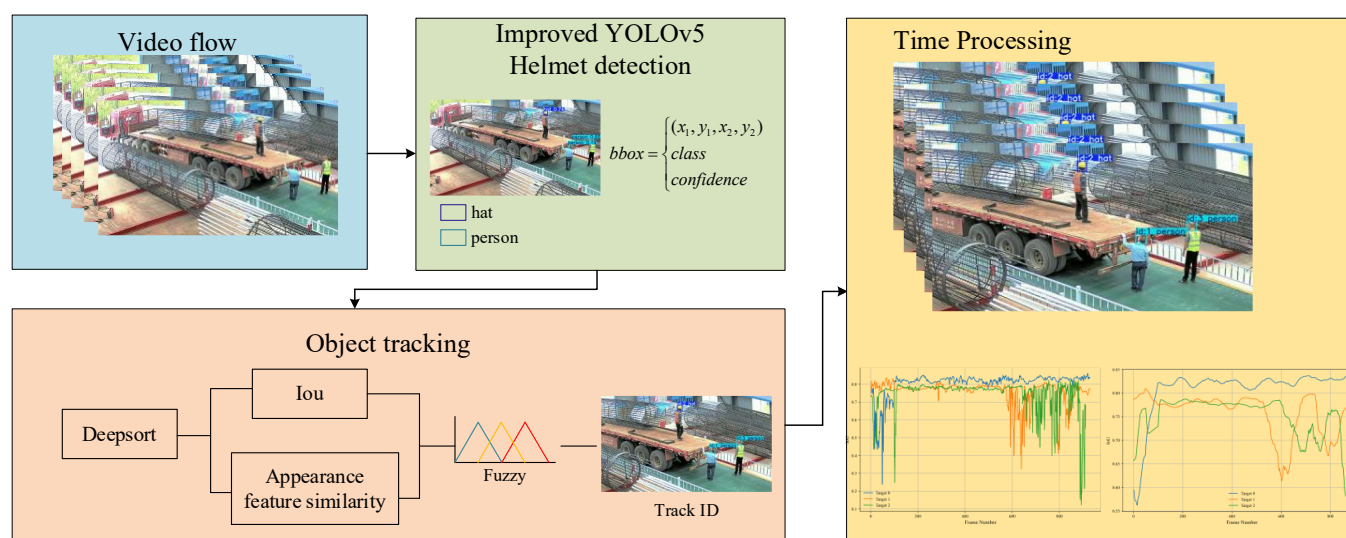


Figure 2. Overall framework of the algorithm process.

3.1. Helmet Detection Algorithm Design

3.1.1. YOLOv5 Algorithm Overview

Since its release, the YOLOv5 algorithm has quickly become the industry standard in the field of object detection due to its exceptional speed, accuracy, and ease of use. The algorithm is fully developed in PyTorch and continues to be optimized with support from the community and ecosystem. As an end-to-end object detection method, YOLOv5 directly predicts the location and class information of bounding boxes through regression, enabling fast and efficient object detection [54].

YOLOv5 offers significant advantages over traditional detection algorithms such as Faster R-CNN. Its model structure includes CSPDarknet as the backbone network and employs PANet to improve feature fusion capabilities, thereby enhancing the detection accuracy for small objects. Additionally, YOLOv5 improves the bounding box localization accuracy through an optimized prediction head, achieving excellent performance on several standard datasets (such as COCO and VOC) [55]. Extensive studies have investigated the YOLOv5 algorithm, and its structure is illustrated in Figure 3.

The network structure of YOLOv5 consists of three main parts: the backbone, the neck, and the head. The backbone network uses CSPDarknet, which improves the network's features and computational efficiency through cross-stage partial connections. The neck structure employs PANet to fuse multiscale features, enhancing the detection accuracy. The head is responsible for the final bounding box regression and class prediction. The backbone network of YOLOv5 extracts deep features from images through multi-level convolution and pooling operations. The PANet structure in the neck fuses features from different scales through path aggregation to capture details of objects of various sizes. The head structure generates the final detection results including the class and position of the objects [56,57].

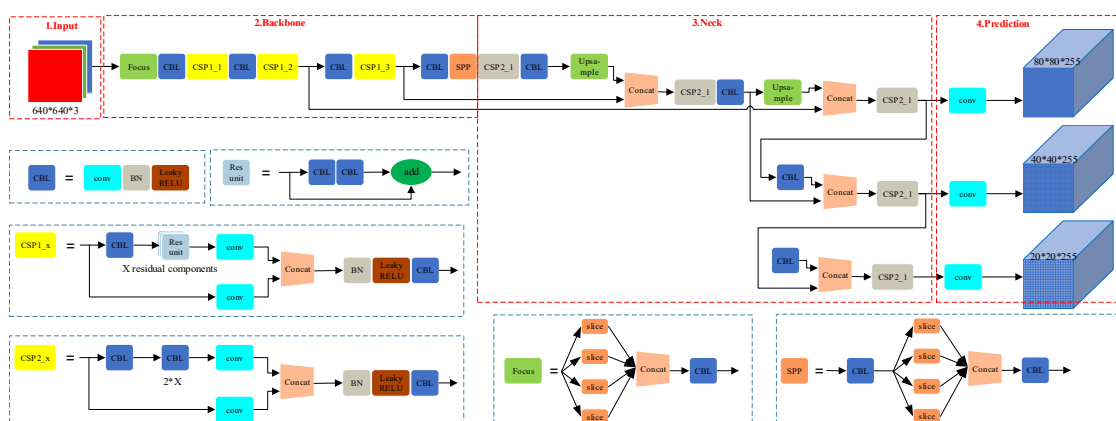


Figure 3. YOLOv5 algorithm network architecture diagram.

Despite the outstanding performance of YOLOv5 in many scenarios, traditional YOLO models still face challenges in certain complex environments, especially in cases of target occlusion and low-resolution images. Therefore, incorporating emerging attention mechanisms to improve YOLOv5's performance, particularly in the detection of small objects such as helmets, is of significant practical importance. To address this, this paper introduces the SACA based on YOLOv5. The goal is to further enhance the model's performance in complex environments through the improved attention mechanism, particularly in improving detection accuracy under occlusion and low contrast scenarios.

3.1.2. SACA

SACA enhances the accuracy of object detection by adaptively adjusting the attention regions in both spatial and channel dimensions, particularly improving the performance in occlusion and complex background scenarios. Unlike traditional attention mechanisms, SACA dynamically calculates the weights for both spatial and channel dimensions and incorporates non-local feature information to better focus the model on key regions. The algorithm flow is shown in Figure 4.

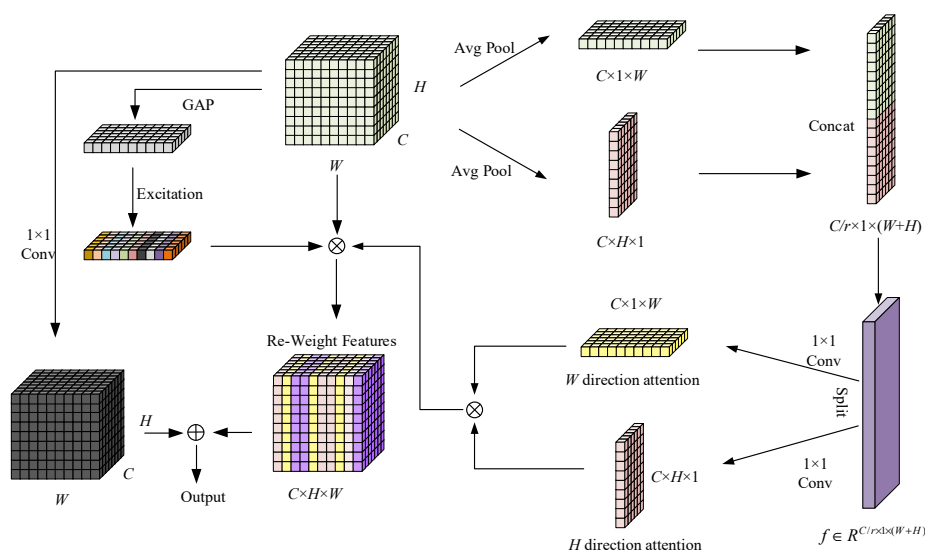


Figure 4. SACA structure diagram.

First, SACA performs adaptive pooling operations on the feature map in both the horizontal and vertical directions, generating feature maps at different scales. Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, horizontal pooling produces $C \times H \times 1$ features and

vertical pooling produces $C \times 1 \times W$ features. After lightweight convolutional operations, both pooled features are broadcast back to the original size $C \times H \times W$ to generate spatial attention weights, ensuring that long-range dependencies in both horizontal and vertical directions are effectively modeled. The pooled feature maps are then fused through convolution operations to generate a spatial attention map. This process is represented as follows:

$$x_h = \text{AdaptiveAvgPool2d}(x) \quad (1)$$

$$x_w = \text{AdaptiveAvgPool2d}(x).\text{permute}(0, 1, 3, 2) \quad (2)$$

$$y = \text{Conv1}([x_h, x_w]) \quad (3)$$

Next, the spatial attention maps a_h and a_w are generated through convolution operations as follows:

$$a_h = \text{sigmoid}(\text{Conv}_h(x_h)) \quad (4)$$

$$a_w = \text{sigmoid}(\text{Conv}_w(x_w)) \quad (5)$$

Then, SACA calculates the channel attention by applying global average pooling and convolution operations to obtain the channel weights as follows:

$$a_c = \text{sigmoid}(\text{Conv}_c(\text{GlobalAvgPool}(x))) \quad (6)$$

To further enhance the perception of global context, SACA introduces a non-local attention mechanism. Traditional local attention typically focuses only on local features, and in scenarios with target occlusion or complex backgrounds, local features often fail to provide sufficient contextual information. To address this issue, SACA integrates global information, enabling the model to capture long-range dependencies, thereby effectively avoiding false detections caused by insufficient local information. Finally, SACA generates the final output feature map by fusing spatial, channel attention, and non-local features as follows:

$$\mathbf{O} = \mathbf{X} \cdot a_h \cdot a_w \cdot a_c + \mathbf{N} \quad (7)$$

where $\mathbf{X}$ is the input feature map, a_h , a_w , and a_c are the spatial, vertical, and channel attention maps, respectively, and $\mathbf{N}$ is the global context feature obtained through non-local attention.

3.2. Multi-Object Tracking Algorithm Design

This paper proposes an improved particle tracking algorithm based on DeepSORT to more accurately track targets in challenging real-world scenarios such as occlusion and rapid motion. The improved DeepSORT structure is shown in Figure 5, with the yellow section representing the improvements. In the original DeepSORT, the algorithm uses a deterministic matching method (IoU matching) and a simple appearance feature extractor. While the algorithm performs well in many scenarios, it struggles in cases of occlusion, fast motion, target re-identification, and unreliable detection. These limitations lead to frequent tracking errors in high-dynamic environments such as incorrect labeling and target loss.

To address these challenges, we introduced a fuzzy logic-based optimization method into the DeepSORT framework. Specifically, we integrated a fuzzy inference system into the target matching process to improve the robustness against occlusion and motion variations. The fuzzy logic component not only uses traditional geometric features (IoU), but also incorporates appearance similarity to evaluate the matching degree between targets. It adjusts the matching process based on uncertain or ambiguous situations. By introducing fuzzy rules, the system can more flexibly adapt to uncertain conditions, maintaining the identity of targets even when their appearance changes or they are partially occluded.

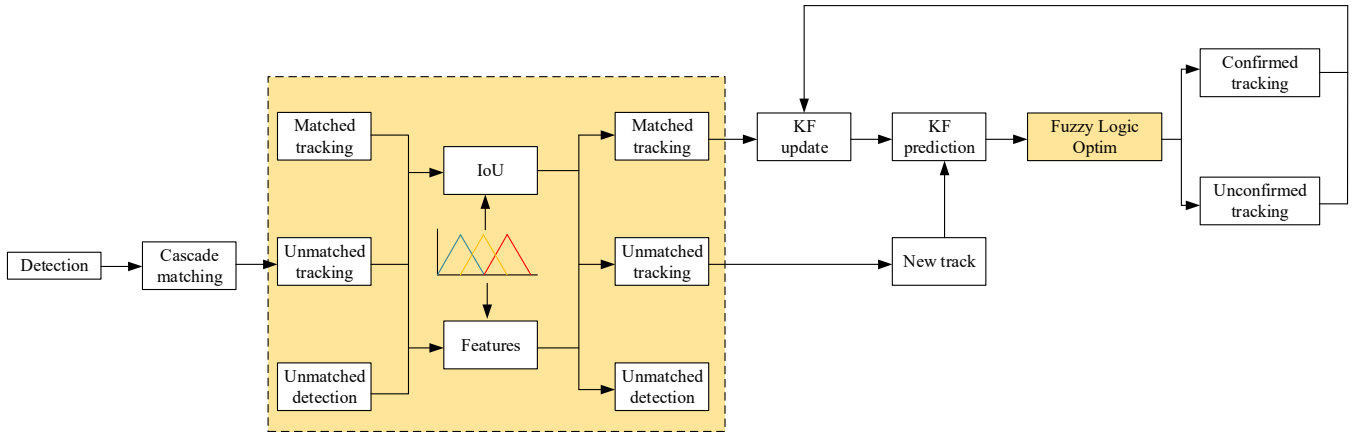


Figure 5. DeepSORT object tracking algorithm with fuzzy logic integration.

The fuzzy logic-based improved structure is as follows:

(a) Fuzzy IoU matching

In the original DeepSORT, the IoU matching threshold is fixed. This method can lead to incorrect matches, particularly when the overlap between targets is small, but their appearance features are very similar. To improve this, we introduced fuzzy logic to dynamically adjust the IoU matching threshold.

We used fuzzy inference to adjust the IoU matching threshold. Through the following formula, we combined the IoU value and appearance similarity to determine the flexibility of the match:

$$\mu_{\text{match}}(\text{IoU}, S) = \mu_{\text{IoU}}(\text{IoU}) \cdot \mu_{\text{appearance}}(S) \quad (8)$$

where $\mu_{\text{IoU}}(\text{IoU})$ is the fuzzy membership degree calculated based on the IoU value, and $\mu_{\text{appearance}}(S)$ is the fuzzy membership degree calculated based on the appearance similarity of the target. IoU is the intersection over union ratio between the bounding boxes, and S is the similarity computed based on appearance features.

This formula dynamically adjusts the matching degree between targets using fuzzy logic, so that when the IoU value is low, if the appearance features are very similar, the targets can still be considered the same. Conversely, if the appearance features differ significantly, despite a high IoU, the targets may be considered different.

The fuzzy membership function for IoU is as follows:

$$\mu_{\text{IoU}}(\text{IoU}) = \begin{cases} 1 & \text{IoU} > 0.8 \\ 0.7 & 0.5 \leq \text{IoU} \leq 0.8 \\ 0.3 & \text{IoU} < 0.5 \end{cases} \quad (9)$$

The fuzzy membership function for appearance similarity is as follows:

$$\mu_{\text{appearance}}(S) = \begin{cases} 1 & S > 0.9 \\ 0.7 & 0.6 \leq S \leq 0.9 \\ 0.4 & S < 0.6 \end{cases} \quad (10)$$

The specific thresholds in Equations (9) and (10) (e.g., IoU values of 0.5 and 0.8, and similarity values of 0.6 and 0.9) were selected based on the distribution analysis of the validation dataset, where they represent transition points between reliable and ambiguous matches. These choices are also consistent with prior studies [52], and can be flexibly adjusted for different application scenarios to balance between stability and sensitivity.

(b) Appearance similarity based on fuzzy inference

Traditional DeepSORT uses the appearance features of targets for matching, but relying solely on the direct comparison of appearance features may not be sufficient to cope with variations in real-world environments (e.g., lighting changes, occlusion, etc.). To address this, we introduced a fuzzy inference mechanism that evaluates appearance similarity through fuzzy rules.

By applying fuzzy logic, the tracking algorithm can better adapt to the complexities of the real-world, such as lighting changes and occlusion, thereby improving the performance and stability in dynamic environments.

We used cosine similarity to calculate the similarity between the appearance features of two targets:

$$S = (\mathbf{f}_1 \cdot \mathbf{f}_2) / (\|\mathbf{f}_1\| \|\mathbf{f}_2\|) \quad (11)$$

where $\mathbf{f}_1$ and $\mathbf{f}_2$ are the appearance feature vectors of target 1 and target 2, and $\|\mathbf{f}\|$ is the L2 norm of the vector.

By introducing fuzzy logic, the improved DeepSORT algorithm becomes more robust during the target matching process, enabling it to better handle complex situations in real-world environments such as lighting changes, occlusion, and target overlap. The fuzzy IoU matching and appearance similarity based on fuzzy inference make the object tracking more flexible and accurate, thereby enhancing the tracking performance and stability in dynamic environments.

3.3. Time Processing Method

To ensure the continuity and stability of helmet wearing status, we introduced a time-based processing method that integrates historical data from the previous frame. This method effectively corrects the issues of target loss and false detection caused by brief occlusion or misdetections, significantly improving the accuracy and stability of helmet wearing status. We recorded the helmet wearing status (HW status) for each frame and marked any lost frames as “NAN”. For false detections or targets temporarily lost, a sliding window was used to process the time-series data. If the number of frames in which the target appears is fewer than the set threshold, it is considered a false detection, and the target’s track is discarded.

The HW status is determined based on the comprehensive matching degree of IoU fuzzy membership and appearance similarity:

$$\text{HW state} = \begin{cases} 0 & \text{IoU} > \text{Threshold} \\ 1 & \text{IoU} \leq \text{Threshold} \end{cases} \quad (12)$$

By combining data from the set frames with information from the current frame, noise can be effectively filtered, and detection accuracy is improved, significantly enhancing the model’s stability in complex dynamic environments.

The size of the sliding window is a critical factor: smaller windows enable faster responsiveness but are more sensitive to noise, while larger windows provide more stable results by reducing false detections at the cost of increased latency. Therefore, the window size should be chosen according to the video frame rate and the complexity of the construction site environment, with a moderate window size offering a good balance between responsiveness and stability.

4. Dataset Description and Experimental Setup

4.1. Helmet Detection Dataset

This paper was primarily divided into two major parts: helmet detection and human tracking. Since the proposed algorithm is a general helmet detection method applicable

to various scenarios, it is not limited to construction environments. Therefore, there is no need to specifically build a large dataset for construction site scenes. Additionally, the algorithm designed in this paper is also suitable for different application scenarios such as traffic and indoor work. As a result, we chose the publicly available SHWD dataset [58] for the experimental dataset.

The SHWD dataset contains 7581 images, divided into two categories: images of workers wearing helmets and those without helmets, focusing on detecting the head region rather than the full body. The images in the dataset were sourced from Google and Baidu, covering a variety of scenarios including traffic, indoor work, and different types of construction environments. These images also include various lighting conditions (e.g., strong dim lighting) and helmets of different colors and shapes, fully demonstrating the method's broad applicability. Figure 6 shows some example images from the dataset.



Figure 6. Examples from the SHWD dataset.

In addition to SHWD, three real construction site surveillance videos were further employed for validation, which contained challenges such as worker occlusion, posture changes, and new worker entries, thereby ensuring that the proposed method was also evaluated under complex real-world conditions. The videos is shown in Section 4.2.

4.2. Helmet-Wearing Worker Tracking

To evaluate the accuracy of our algorithm, we selected three sets of real worker site monitoring video datasets, named Video1, Video2, and Video3. Figure 7 shows the cropped images from these three videos over time. The following is a detailed description of the content of these three videos. All videos were filmed using the same Hikvision camera, with a frame rate of 20 and a resolution of 1920×1080 .

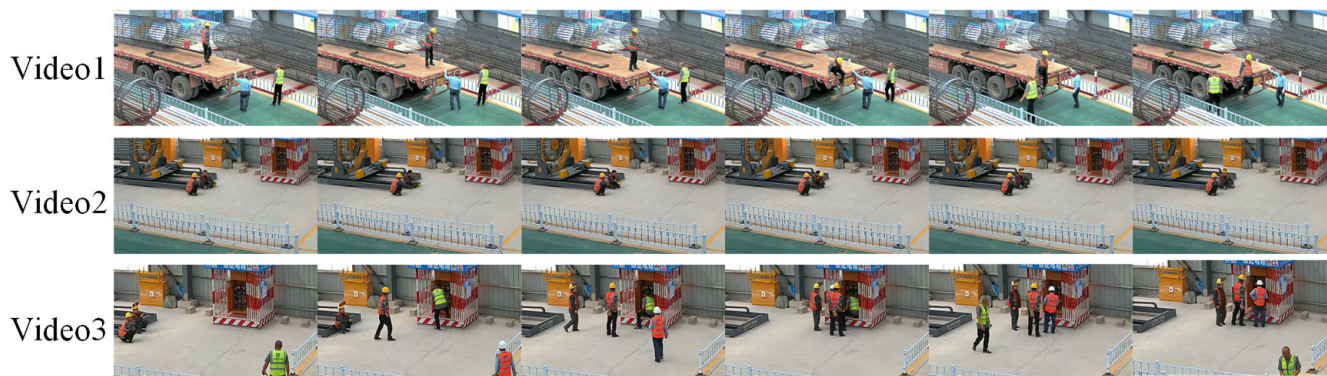


Figure 7. Example frames from Video1, Video2, and Video3.

Video1: This video shows a worker wearing a helmet standing on a truck, while two other workers are working on the ground. The worker on the truck performs an exit action, showing a significant posture change, from sitting down to jumping off. Meanwhile, the two workers on the ground are in walking and standing positions, with one worker experiencing a brief occlusion during their walking process.

Video2: This video shows two workers sitting on the ground. One worker is wearing a helmet, while the other is not, with the helmet placed on the ground. The video clearly records the tracking changes of the same worker before and after wearing the helmet by showing the actions of putting on the helmet.

Video3: This video shows four workers working in front of an electrical distribution box. Initially, two workers wearing helmets are sitting, and they later join in the work at the distribution box. Another worker without a helmet starts working immediately, and the last worker, a new member, joins the operation. This video includes worker occlusion and a new worker joining the scene, increasing the complexity of the tracking task.

4.3. Experimental Setup

The experimental environment for training on the SHWD dataset and target tracking is shown in Table 1.

Table 1. Experimental platform and environment.

Component	Parameters
CPU	13th Gen Intel(R) Core(TM)i7-14700KF
RAM	32 G
GPU	NVIDA GeForce RTX 4070 12 GB
Programing language	Python 3.8
Deep learning framework	PyTorch 2.3.1
CUDA	11.8

After experimental tuning, the training settings were as follows: the number of epochs was set to 50, the batch size was 16, the initial learning rate was set to 0.001, and the IoU threshold was 0.5. The model parameters were optimized using the Adam optimizer.

5. Experimental Results

5.1. Helmet Detection Experimental Results

In this study, precision, recall, and mAP@0.5 were used as the evaluation metrics for helmet detection and multi-object tracking tasks, as described below:

First, precision was used to measure the proportion of true positive samples among all the samples classified as positive by the model. The formula for calculating precision is:

$$\text{Precision} = TP / (TP + FP) \quad (13)$$

where TP represents the number of true positives and FP represents the number of false positives. A higher precision indicates a lower probability of misclassification when detecting positive class targets.

Next, recall was used to measure the proportion of correctly identified positive samples among all of the actual positive samples. The formula for calculating recall is:

$$\text{Recall} = TP / (TP + FN) \quad (14)$$

where FN represents the number of false negatives. A higher recall indicates that the model is able to detect more positive class targets, but it may also result in more false positives.

To comprehensively evaluate the model's detection performance, we used mAP@0.5 as the evaluation metric. mAP@0.5 is calculated based on the area under the precision–recall curve, where the IoU threshold was set to 0.5. The formula for calculating mAP@0.5 is:

$$\text{mAP@0.5} = \frac{1}{N} \sum_{i=1}^N AP_i \quad (15)$$

where N represents the number of categories, and AP_i is the average precision for the i -th class. The average precision was obtained by calculating the area under the precision–recall curve. A higher mAP indicates that the model performs better across different scenarios, especially in terms of robustness and accuracy in complex environments.

In this paper, the results obtained from training the proposed improved YOLOv5 algorithm on the SHWD dataset are shown in Figure 7. Figure 8a displays all the ground truth bounding boxes, while Figure 8b,c show that the height and width of the ground truth boxes were smaller than the height and width of the image, demonstrating the validity of the training.

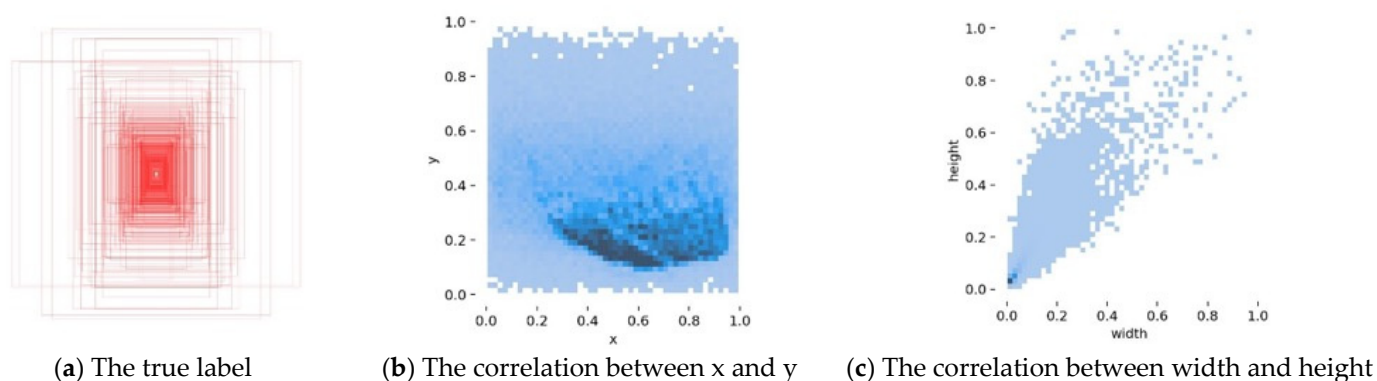


Figure 8. Internal structure of the SHWD dataset.

Figure 9 shows the PR curve obtained from the improved YOLOv5 model we designed. As can be seen, the model achieved 91.6% accuracy for helmet detection, 96.3% accuracy for non-helmet detection, and an overall accuracy of 94.0%. This indicates that the model is particularly well-suited for detecting unsafe behaviors related to the absence of helmets.

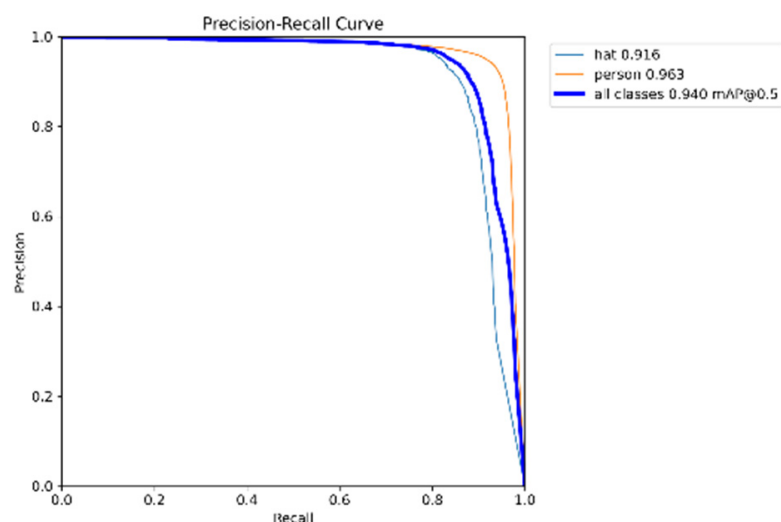


Figure 9. PR curve of model training.

To comprehensively evaluate the performance of the proposed model, it was compared not only with several recent YOLO algorithms, but also with typical attention-based variants of YOLOv5 such as SE, CBAM, and ECA. All algorithms were trained for 50 epochs, with a batch size of 16, to ensure fairness in the training conditions. The results are shown in Table 2. It can be observed that the proposed method performed excellently on the mAP@0.5 metric, achieving a value of 0.940, placing it in a leading position. Although the precision metric was not the highest, the model's computational load had nearly not increased, and the number of parameters had not significantly grown based on the improvements made to YOLOv5. Maintaining such excellent performance represents a significant advancement. In addition, compared with the SE, CBAM, and ECA attention modules, the proposed SACA achieved a higher mAP and recall, showing its stronger capability in handling small targets and occlusion in construction site scenarios. This result fully validates the effectiveness of the introduced SACA, demonstrating that it can improve the detection accuracy while effectively controlling the computational resource consumption.

Table 2. Comparison of indicators of helmet detection by different YOLO algorithms.

Model	Precision	Recall	mAP@0.5
YOLOv5	0.931	0.849	0.907
YOLOv7	0.935	0.852	0.912
YOLOv8	0.933	0.857	0.919
YOLOv9	0.934	0.864	0.924
YOLOv5 + SE	0.936	0.872	0.926
YOLOv5 + CBAM	0.938	0.878	0.931
YOLOv5 + ECA	0.937	0.874	0.928
Improved YOLOv5	0.924	0.890	0.940

Figure 10 presents a comparison of the three metrics for our improved YOLOv5 algorithm and several typical algorithms from Table 2. From the results, it can be seen that while our algorithm did not show a significant advantage in precision, it consistently led in both recall and mAP@0.5. This indicates that our model was more effective at correctly identifying instances.

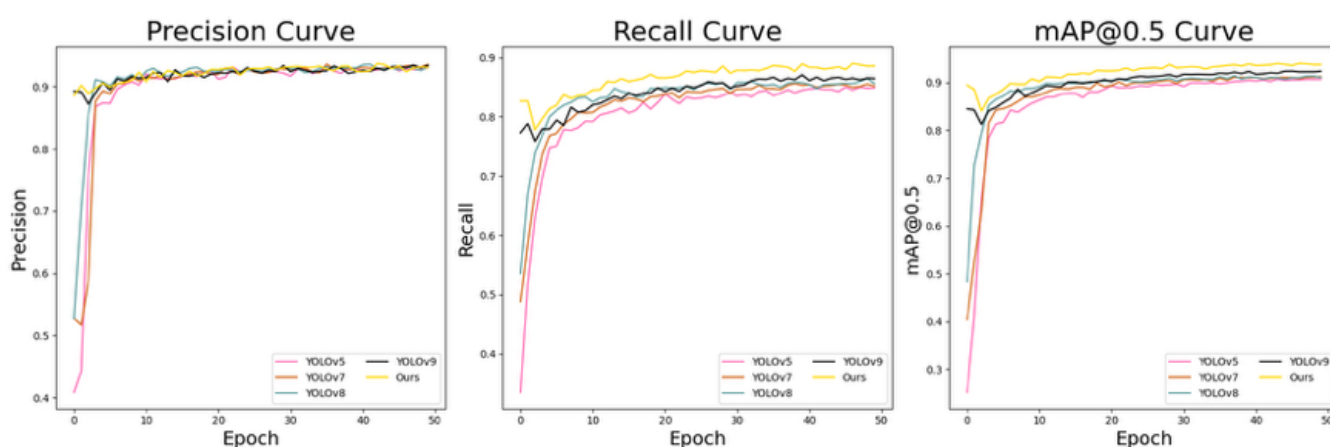


Figure 10. Comparison of three metrics for various algorithms, from left to right: precision, recall, and mAP@0.5.

In addition to accuracy, efficiency metrics were also evaluated to verify the real-time applicability of the proposed method. Table 3 summarizes the comparison of training time, inference speed (FPS), and GPU utilization with the baseline YOLOv5s and lightweight variants. The results showed that the proposed YOLOv5s + SACA achieved 28 FPS on an

NVIDIA RTX 4070 GPU, which is sufficient for real-time helmet detection in construction sites, while maintaining higher accuracy than the lightweight alternatives.

Table 3. Comparison of the accuracy and efficiency metrics (training time, FPS, GPU usage) across models.

Model	Params (M)	FLOPs (G)	Inference FPS (RTX 4070)	mAP@0.5
YOLOv5s (baseline)	7.2	16.5	34	0.907
YOLOv5s + CBAM	7.8	17.9	30	0.931
YOLOv5s + SACA	8.0	18.2	28	0.940
YOLOv8n	3.2	8.7	40	0.919

5.2. Multi-Object Tracking for Helmet Detection

Figure 11 shows the tracking results for the targets in Video1 using the proposed DeepSORT tracking algorithm with fuzzy logic integration. As previously mentioned, Video1 primarily demonstrates the detection of worker motion changes and occlusion phenomena. From Figure 10, it can be seen that for helmet detection, the IoU matching score remained consistently high, indicating that the method effectively tracked helmets in Video1.

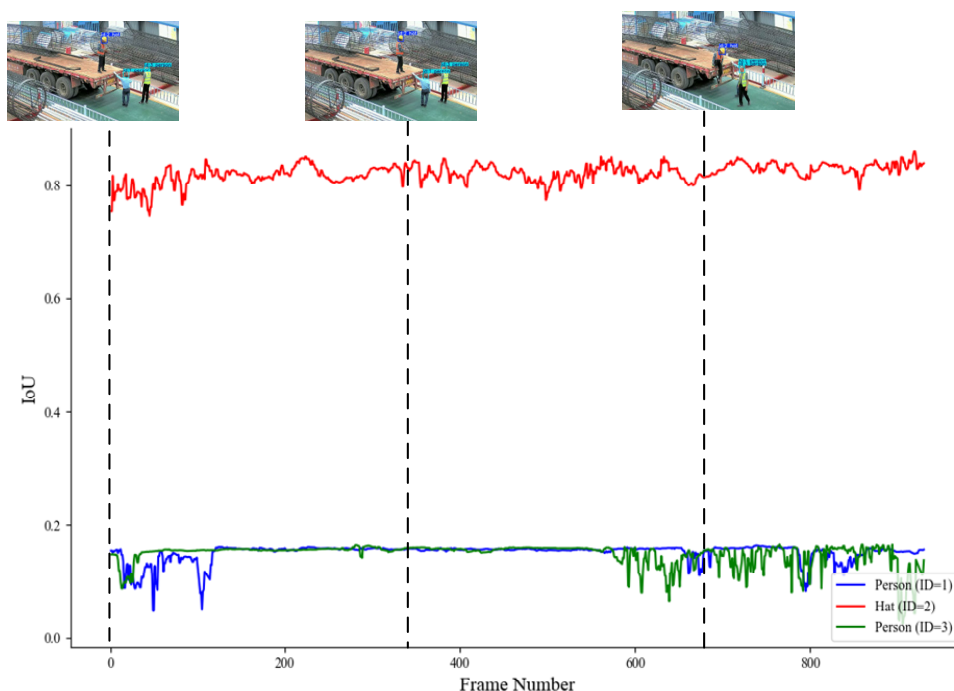


Figure 11. IoU matching in Video1.

Although in person detection the IoU matching score was lower during some periods due to occlusion of the worker, the overall tracking still fluctuated, and no misidentifications occurred. This suggests that the model can stably track whether workers are wearing helmets and demonstrates excellent robustness in complex scenarios.

In Video2, the process of a worker transitioning from not wearing a helmet to wearing one is demonstrated. Figure 12 compares the tracking performance between the traditional DeepSORT and the optimized DeepSORT on sample frames. It can be observed that the traditional DeepSORT encounters issues of missed helmet detection and ID switching errors when the worker puts on the helmet.

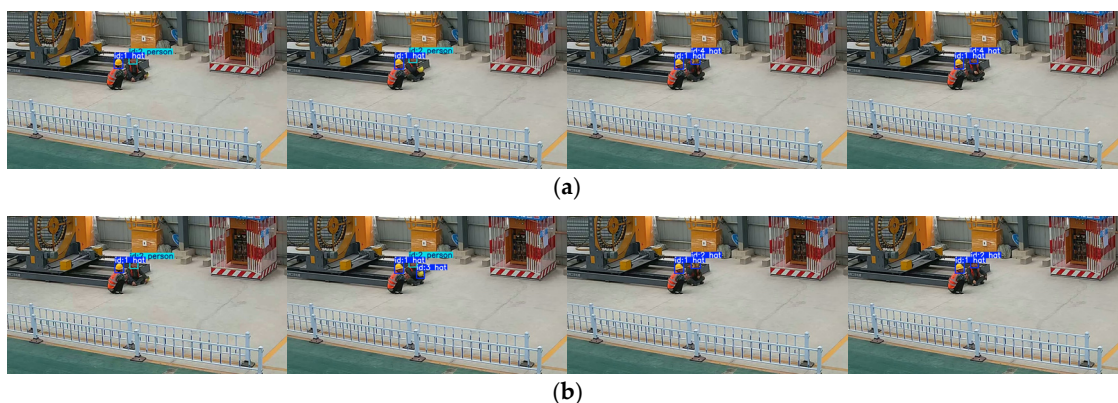


Figure 12. (a) Traditional DeepSORT tracking of helmet-wearing status in Video2. (b) Optimized DeepSORT tracking of helmet-wearing status in Video2.

In contrast, the optimized DeepSORT, which incorporates fuzzy logic and appearance similarity, not only avoided the issue of missed helmet detection, but also accurately identified the changes in the helmet-wearing status of the same worker by introducing a temporal processing algorithm. This result fully demonstrates the superior performance and robustness of the optimized algorithm in complex scenarios.

To ensure temporal stability of helmet-wearing status across frames, a sliding-window mechanism was adopted to process the time-series data. The window size directly affects the trade-off between temporal smoothness and responsiveness. A smaller window can quickly reflect status transitions but is more susceptible to transient detection noise, while a larger window provides smoother predictions but introduces latency.

Through empirical experiments on 20 fps construction-site videos, we evaluated window sizes from 5 to 50 frames. As shown in Figure 13, a 20–30 frame window yielded the best balance, effectively filtering short-term noise while maintaining prompt responses to real changes. Hence, the optimal window length can be adapted according to the frame rate and the temporal dynamics of the specific application scenario.

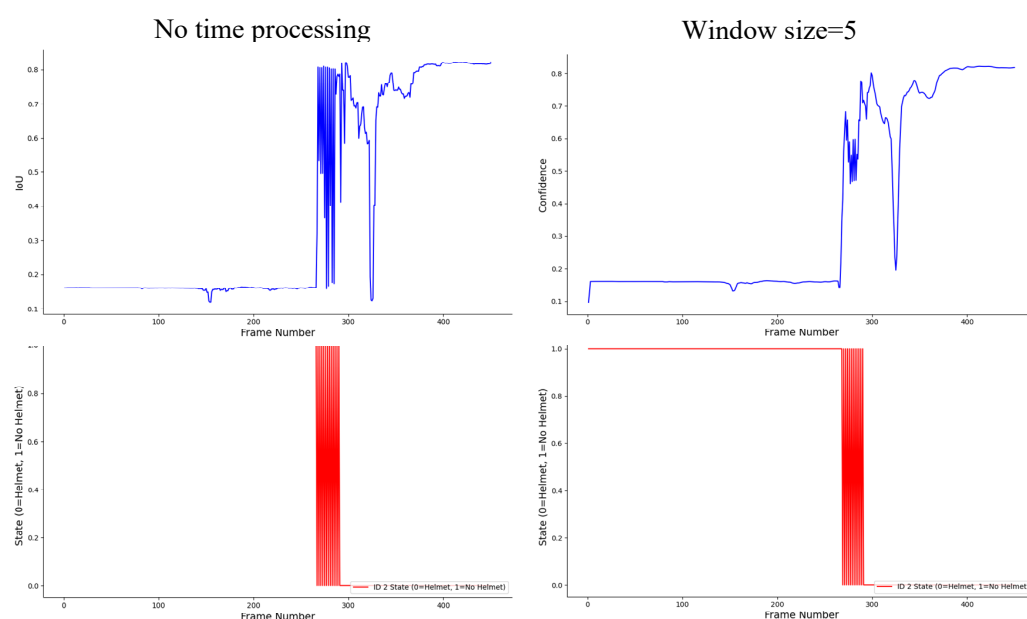


Figure 13. Cont.

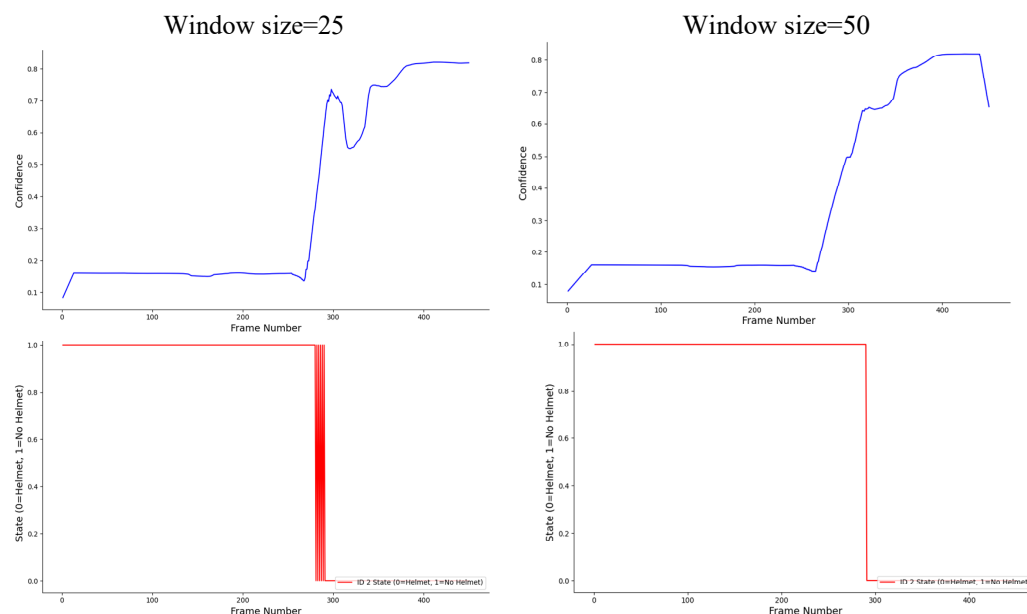


Figure 13. Comparison of temporal processing for helmet-wearing status detection of the second worker in Video2.

From Figure 11, it is evident that without temporal processing, incorrect judgments occurred, especially during the process of wearing the helmet, where misidentifications repeatedly took place. However, after smoothing with a window function, this issue was significantly alleviated. When the window size was set to 50, false detections were almost entirely eliminated. Nevertheless, the window size must be carefully chosen; a window that is too large may lead to data loss, which could affect the accuracy of the results.

For Video3, which was relatively more complex and includes long periods of worker occlusion as well as the introduction of other workers, the video effectively validates the tracking performance of the proposed algorithm in long-duration videos. Figure 14a,b show examples of multi-worker tracking using the traditional DeepSORT algorithm and the optimized DeepSORT algorithm, respectively. From Figure 14, it can be seen that the traditional DeepSORT algorithm encountered ID mismatches when new workers appeared and failed to recover the original worker's ID during long occlusion periods. In contrast, the optimized DeepSORT algorithm, although still experiencing brief ID loss, was able to accurately recover the original worker's ID after a short time, fully demonstrating the stability and robustness of the proposed algorithm in long-duration videos.

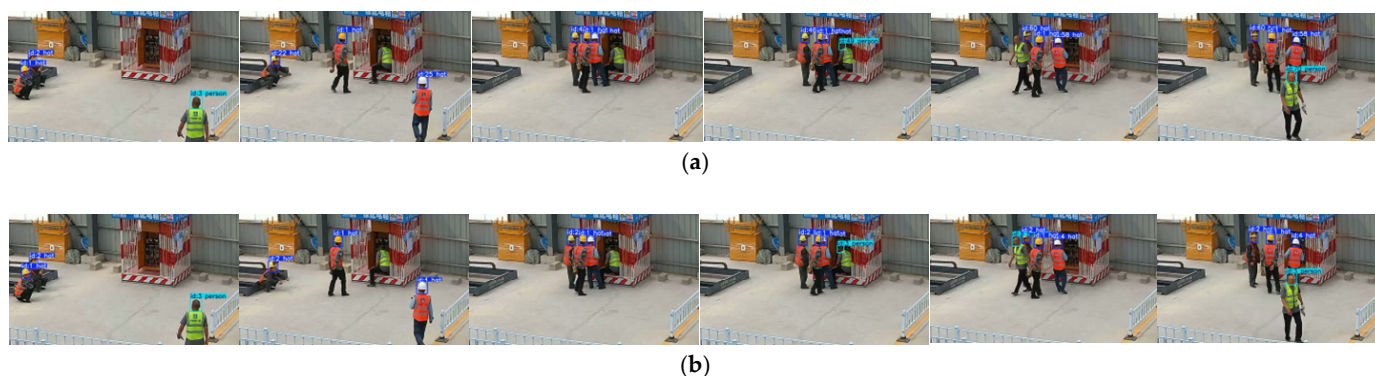


Figure 14. (a) Traditional DeepSORT tracking of helmet-wearing status in Video3. (b) Optimized DeepSORT tracking of helmet-wearing status in Video3.

For multi-object tracking, we used the following three evaluation metrics to measure the overall performance of the tracking algorithm.

(a) MOTA (multiple object tracking accuracy)

MOTA takes into account false positives (FP), false negatives (FN), and identity switches (IDSw), and is calculated using the following formula:

$$\text{MOTA} = 1 - (FN + FP + \text{IDSw}) / GT \quad (16)$$

where GT represents the total number of ground truth objects. A higher MOTA value indicates a lower probability of errors during tracking, signifying more accurate and stable tracking results.

(b) IDF1 (Identity F1 Score)

IDF1 is a metric that measures the performance of target identity recognition, combining the number of correctly identified target identities with the number of mismatches. The formula for calculating IDF1 is

$$\text{IDF1} = \frac{2 \times \text{IDTP}}{2 \times \text{IDTP} + \text{IDFP} + \text{IDFN}} \quad (17)$$

where IDTP represents the number of correctly matched target identities, IDFP represents the number of incorrectly matched target identities, and IDFN represents the number of unmatched target identities.

(c) IDS (Identity Switches)

IDS refers to the number of times the model incorrectly switches the identity of a target during multi-object tracking. Identity switches cause instability in tracking, so a lower IDS value indicates better performance in maintaining target identity consistency. This metric does not have a fixed formula and is evaluated by calculating the number of target identity switches.

The comparison of the proposed algorithm with several advanced methods is shown in Tables 4 and 5. To ensure fairness and better isolate the contributions of detection backbones and tracking strategies, we report the results under two settings: (i) fixing the detector (YOLOv5-SACA) while varying the tracker, and (ii) fixing the tracker (DeepSORT with fuzzy logic and temporal processing) while varying the detector. In addition, we included several representative models such as anchor-free detectors (CenterNet), transformer-based detectors (Deformable DETR), and joint detection–tracking methods (FairMOT).

Table 4. Tracking performance comparison with YOLOv5-SACA as the detector.

Tracker	MOTA	IDF1	IDS
StrongSORT	85.2%	80.3%	12
BoT-SORT	88.7%	83.6%	10
ByteTrack	91.0%	85.1%	9
OC-SORT	87.2%	82.0%	11
FairMOT (joint MOT)	86.1%	81.5%	14
DeepSORT	86.3%	82.4%	8
DeepSORT + Fuzzy + Time processing	90.5%	84.2%	5

From Table 4, where YOLOv5-SACA was fixed as the detector, different tracking algorithms achieved varying levels of performance. ByteTrack achieved the highest MOTA and IDF1 values (91.0% and 85.1%), but at the cost of increased computational complexity and GPU memory consumption, which may limit its deployment in resource-constrained construction environments. In contrast, the proposed DeepSORT with fuzzy logic and temporal processing achieved a slightly lower accuracy, but with the lowest number of

identity switches (5), indicating more stable long-term tracking and a better trade-off between accuracy and efficiency.

Table 5. Tracking performance comparison with DeepSORT + fuzzy logic as the tracker.

Detector	MOTA	IDF1	IDS
YOLOv5 (baseline)	85.1%	81.0%	11
YOLOv8	88.2%	83.5%	9
YOLOv9	91.3%	85.6%	8
CenterNet (anchor-free)	87.0%	82.2%	12
Deformable DETR (transformer)	90.1%	84.5%	10
YOLOv5-SACA (proposed)	90.5%	84.2%	5

From Table 5, where DeepSORT with fuzzy logic was fixed as the tracker, YOLOv9 and Deformable DETR achieved strong detection performance, slightly surpassing the proposed YOLOv5-SACA in MOTA and IDF1. However, YOLOv9 requires significantly more parameters and FLOPs, while transformer-based models like Deformable DETR introduce high inference latency, which restricts their real-time applicability. In contrast, the proposed YOLOv5-SACA achieved comparable accuracy with lower computational overhead and the lowest IDS value, making it more suitable for real-time helmet detection in complex construction site scenarios.

6. Conclusions and Discussion

6.1. Conclusions

This study explored the application of the multi-object tracking method based on YOLOv5 and DeepSORT in construction site helmet detection. This research found that combining the YOLOv5 detection model with the DeepSORT tracking algorithm demonstrated high accuracy in monitoring helmet-wearing status. By introducing the SACA and temporal processing algorithm, the model's stability and robustness in complex scenarios were enhanced. Experimental results showed that the proposed model outperformed existing multi-object tracking algorithms in terms of the MOTA, IDF1, and IDS metrics, with significant advantages, especially in long-duration videos and occlusion scenarios. In particular, the introduction of the SACA further enhanced the detection accuracy, proving its effectiveness in handling dynamic worker state changes.

However, there are still some limitations in the proposed method. Although the introduction of the temporal processing algorithm reduced misdetections and identity switches when tracking multiple workers, there may still be occasional loss and misjudgment in extreme occlusion or complex backgrounds. Additionally, while the SACA mechanism significantly improved the detection accuracy, its computational overhead also increased. Future work could focus on further optimizing the computational efficiency, especially for applications on large-scale datasets. Future research may also concentrate on combining temporal processing with other deep learning methods (e.g., Transformer models) to further enhance detection and tracking performance in multi-worker scenarios.

6.2. Discussion

The experimental results demonstrate that the proposed method achieved excellent performance in construction site helmet detection tasks, particularly in long-duration videos and occlusion scenarios, showing a clear advantage over existing algorithms. The proposed approach integrates the detection capability of YOLOv5 with the DeepSORT tracking algorithm, fully leveraging the strengths of deep learning in both object detection and multi-object tracking. By incorporating the self-adaptive coordinate attention (SACA) mechanism,

the model effectively captured subtle variations in workers' helmet-wearing behavior, thereby improving the detection accuracy. Furthermore, the temporal processing algorithm enhanced the model stability when handling long video sequences, reducing misdetections and identity switches. In addition to the SHWD dataset, the proposed method was also validated on three real construction site surveillance videos, demonstrating its effectiveness and robustness in complex, real-world environments. As large-scale public construction site datasets remain limited, we supplemented SHWD with self-collected footage for training and validation. Future work will focus on constructing larger and more diverse datasets to enable more comprehensive evaluation and further generalization analysis.

However, despite the significant progress achieved in multi-object tracking, there are still some challenges. First, detection under extreme occlusion remains a difficult problem, especially when multiple targets overlap or move rapidly. In such cases, the detector may fail to generate reliable bounding boxes, causing both IoU and appearance similarity features to become unreliable. Meanwhile, the Kalman filter predictions may drift during long-term occlusion, and once the target reappears, it is sometimes assigned a new identity, leading to false detections and ID switches. Second, while the SACA mechanism effectively improves the detection accuracy, it also increases the computational complexity of the model. Future research could therefore explore more efficient attention mechanisms or optimization strategies to reduce the computational burden. Finally, this method primarily relies on static video frames for detection and tracking. Future research could investigate video-level detection frameworks or temporal modeling approaches (e.g., Transformer-based methods) to further enhance the model's performance in real-world applications.

In future work, deep learning-based multi-object tracking methods can be further integrated with self-supervised and semi-supervised learning techniques to reduce dependence on large-scale labeled datasets. In construction site environments, where workers' helmet-wearing behaviors are highly dynamic, employing self-supervised learning for model pre-training can effectively enhance the adaptability and generalization across diverse scenarios. Moreover, for cross-frame target tracking, future research could incorporate richer temporal information and reinforcement learning strategies to further improve the accuracy and stability of detection and tracking. With these advancements, helmet detection systems can be more effectively applied to real-world construction environments, contributing to more efficient and reliable safety management systems, ultimately strengthening worker protection.

Author Contributions: Conceptualization, X.Z., X.J., and J.B.; Methodology, X.Z., and X.L. (Xiang Lv); Software, X.Z., and X.L. (Xiaodong Lv); Validation, X.J., J.B., and X.L. (Xiang Lv); Formal analysis, X.Z., J.B., and X.L. (Xiaodong Lv); Investigation, X.Z.; Resources, G.Z., and X.J.; Data curation, X.Z., X.J., and J.B.; Writing—original draft preparation, X.Z., and X.L. (Xiaodong Lv); Writing—review and editing, X.J., and X.L. (Xiang Lv); Visualization, X.Z., J.B., and X.L. (Xiaodong Lv); Supervision, G.Z.; Project administration, G.Z.; Funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China, grant number 1020241708, the China Civil Engineering Society Research Project (CCES2024KT10), the China State Construction Engineering Corporation Technology R&D Program (CSCEC-2024-Q-74), the Chongqing Construction Science and Technology Program (CQCS2024-8-5), and the Jiangsu Province Science and Technology Program (No. BX2023).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: Xuejun Jia, Jian Bai and Xiang Lv are employed by China Construction Second Engineering Bureau Co., Ltd. They declare that there is no conflict of interest. The remaining authors declare that there is no conflict of interest.

References

- Halabi, Y.; Xu, H.; Long, D.B.; Chen, Y.H.; Yu, Z.X.; Alhaek, F.; Alhaddad, W. Causal factors and risk assessment of fall accidents in the US construction industry: A comprehensive data analysis (2000–2020). *Saf. Sci.* **2022**, *146*, 105537. [\[CrossRef\]](#)
- Cheng, M.Y.; Soegiono, D.V.N.; Khitam, A.F.K. Automated fall risk classification for construction workers using wearable devices, BIM, and optimized hybrid deep learning. *Autom. Constr.* **2025**, *172*, 106072. [\[CrossRef\]](#)
- Liang, H.; Seo, S. UAV Low-Altitude Remote Sensing Inspection System Using a Small Target Detection Network for Helmet Wear Detection. *Remote. Sens.* **2023**, *15*, 196. [\[CrossRef\]](#)
- Song, R.J.; Wang, Z.M. RBFPDet: An anchor-free helmet wearing detection method. *Appl. Intell.* **2022**, *53*, 5013–5028. [\[CrossRef\]](#)
- Jing, W.Y.; Zhu, Z.J.; Chen, H.W.; Wang, H.Z.; Shao, F. Toward Large-Scale Non-Motorized Vehicle Helmet Wearing Detection: A New Benchmark and Beyond. *IEEE Trans. Consum. Electron.* **2025**, *71*, 594–607. [\[CrossRef\]](#)
- Lee, J.Y.; Choi, W.S.; Choi, S.H. Verification and performance comparison of CNN-based algorithms for two-step helmet-wearing detection. *Expert Syst. Appl.* **2023**, *225*, 120096. [\[CrossRef\]](#)
- Ren, H.G.; Fan, A.N.; Zhao, J.; Song, H.R.; Wen, Z.G.; Lu, S. A dynamic weighted feature fusion lightweight algorithm for safety helmet detection based on YOLOv8. *Measurement* **2025**, *253*, 117572. [\[CrossRef\]](#)
- Li, J.P.; Yang, B.S.; Yang, Y.D.; Zhao, X.; Liao, Y.Q.; Zhu, N.N.; Dai, W.; Liu, R.; Chen, R.; Dong, Z. Real-time automated forest field inventory using a compact low-cost helmet-based laser scanning system. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *118*, 103299. [\[CrossRef\]](#)
- Zhang, Y.; Huang, S.Z.; Qin, J.Z.; Li, X.Y.; Zhang, Z.X.; Fan, Q.H.; Tan, Q. Detection of helmet use among construction workers via helmet-head region matching and state tracking. *Autom. Constr.* **2025**, *171*, 105987. [\[CrossRef\]](#)
- Shi, Y.M.; Du, J.; Ahn, C.R.; Ragan, E. Impact assessment of reinforced learning methods on construction workers' fall risk behavior using virtual reality. *Autom. Constr.* **2019**, *104*, 197–214. [\[CrossRef\]](#)
- Peir, G.H. Slips, trips, and falls: A quality improvement initiative. *Pediatr. Qual. Saf.* **2022**, *7*, e550. [\[CrossRef\]](#)
- Wan, H.P.; Zhang, W.J.; Ge, H.B.; Luo, Y.Z.; Todd, M.D. Improved Vision-Based Method for Detection of Unauthorized Intrusion by Construction Sites Workers. *J. Constr. Eng. Manag.* **2023**, *149*, 04023040. [\[CrossRef\]](#)
- Lee, Y.R.; Jung, S.H.; Kang, K.S.; Ryu, H.C.; Ryu, H.G. Deep learning-based framework for monitoring wearing personal protective equipment on construction sites. *J. Comput. Des. Eng.* **2023**, *10*, 905–917. [\[CrossRef\]](#)
- Mei, X.Y.; Xu, F.; Zhang, Z.P.; Tao, Y. Unsafe behavior identification on construction sites by combining computer vision and knowledge graph-based reasoning. *Eng. Constr. Arch. Manag.* **2024**, ahead of print. [\[CrossRef\]](#)
- Kim, K.; Kim, K.; Jeong, S. Application of YOLO v5 and v8 for Recognition of Safety Risk Factors at Construction Sites. *Sustainability* **2023**, *15*, 15179. [\[CrossRef\]](#)
- Wu, H.; Zhao, J. An intelligent vision-based approach for helmet identification for work safety. *Comput. Ind.* **2018**, *100*, 267–277. [\[CrossRef\]](#)
- He, G.; Qi, D. A Keypoint-guided Pipeline for Safety Violation Identification. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 7223–7228.
- Shen, J.; Xiong, X.; Li, Y.; He, W.; Li, P.; Zheng, X. Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning. *Comput. Aided Civ. Infrastruct. Eng.* **2021**, *36*, 180–196. [\[CrossRef\]](#)
- Farooq, M.U.; Bhutto, M.A.; Kazi, A.K. Real-Time Safety Helmet Detection Using YOLOv5 at Construction Sites. *Intell. Autom. Soft Comput.* **2023**, *36*, 911–927. [\[CrossRef\]](#)
- Hayat, A.; Morgado-Dias, F. Deep learning-based automatic safety helmet detection system for construction safety. *Appl. Sci.* **2022**, *12*, 8268. [\[CrossRef\]](#)
- Ngoc-Thoan, N.; Bui, D.-Q.T.; Tran, C.N.; Tran, D.-H. Improved detection network model based on YOLOv5 for warning safety in construction sites. *Int. J. Constr. Manag.* **2024**, *24*, 1007–1017. [\[CrossRef\]](#)
- Cai, J.; Yang, L.; Zhang, Y.; Li, S.; Cai, H. Multitask learning method for detecting the visual focus of attention of construction workers. *J. Constr. Eng. Manag.* **2021**, *147*, 04021063. [\[CrossRef\]](#)
- Zhao, S.X.; Zhong, R.Y.; Jiang, Y.S.; Besklubova, S.; Tao, J.; Yin, L. Hierarchical spatial attention-based cross-scale detection network for Digital Works Supervision System (DWSS). *Comput. Ind. Eng.* **2024**, *192*, 110220. [\[CrossRef\]](#)
- Liang, H.; Cho, J.; Seo, S. Construction Site Multi-Category Target Detection System Based on UAV Low-Altitude Remote Sensing. *Remote. Sens.* **2023**, *15*, 1560. [\[CrossRef\]](#)
- Kim, B.; An, E.J.; Kim, S.; Preethaa, K.R.S.; Lee, D.E.; Lukacs, R.R. SRGAN-enhanced unsafe operation detection and classification of heavy construction machinery using cascade learning. *Artif. Intell. Rev.* **2024**, *57*, 206. [\[CrossRef\]](#)

26. Ma, P.; He, X.Y.; Chen, Y.Y.; Liu, Y. ISOD: Improved small object detection based on extended scale feature pyramid network. *Vis. Comput.* **2024**, *41*, 465–479. [[CrossRef](#)]
27. Wang, M.P.; Yao, G.; Yang, Y.; Sun, Y.J.; Yan, M.; Deng, R. Deep learning-based object detection for visible dust and prevention measures on construction sites. *Dev. Built Environ.* **2023**, *16*, 100245. [[CrossRef](#)]
28. Zhang, Y.F.; Ren, W.Q.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T.N. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [[CrossRef](#)]
29. Jia, X.J.; Zhou, X.X.; Shi, Z.H.; Xu, Q.; Zhang, G.M. GeoIoU-SEA-YOLO: An Advanced Model for Detecting Unsafe Behaviors on Construction Sites. *Sensors* **2025**, *25*, 1238. [[CrossRef](#)]
30. Pisu, A.; Elia, N.; Pompianu, L.; Barchi, F.; Acquaviva, A.; Carta, S. Enhancing workplace safety: A flexible approach for personal protective equipment monitoring. *Expert Syst. Appl.* **2024**, *238*, 122285. [[CrossRef](#)]
31. Ahmed, M.I.B.; Saraireh, L.; Rahman, A.; Al-Qarawi, S.; Mhran, A.; Al-Jalaoud, J.; Al-Mudaifer, D.; Al-Haidar, F.; AlKhulaifi, D.; Youldash, M.; et al. Personal Protective Equipment Detection: A Deep-Learning-Based Sustainable Approach. *Sustainability* **2023**, *15*, 13990. [[CrossRef](#)]
32. Ke, X.; Chen, W.Y.; Guo, W.Z. 100+FPS detector of personal protective equipment for worker safety: A deep learning approach for green edge computing. *Peer Peer Netw. Appl.* **2022**, *15*, 950–972. [[CrossRef](#)]
33. Fang, Q.; Li, H.; Luo, X.C.; Ding, L.Y.; Luo, H.B.; Rose, T.M.; An, W. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* **2018**, *85*, 1–9. [[CrossRef](#)]
34. Li, K.S.; Wang, J.C.; Jalil, H.; Wang, H. A fast and lightweight detection algorithm for passion fruit pests based on improved YOLOv5. *Comput. Electron. Agric.* **2022**, *204*, 107534. [[CrossRef](#)]
35. Li, J.Q.; Zhao, X.F.; Zhou, G.Y.; Zhang, M.Y. Standardized use inspection of workers' personal protective equipment based on deep learning. *Saf. Sci.* **2022**, *150*, 105689. [[CrossRef](#)]
36. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
37. Gu, Y.W.; Wang, Y.S.; Shi, L.; Li, N.; Zhuang, L.H.; Xu, S.K. Automatic detection of safety helmet wearing based on head region location. *Let Image Process.* **2021**, *15*, 2441–2453. [[CrossRef](#)]
38. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [[CrossRef](#)]
39. Chou, J.S.; Chen, C.J.; Liu, C.Y. Amodal instance segmentation optimized by metaheuristics for enhanced safety behavior detection on construction sites. *Autom. Constr.* **2025**, *178*, 106412. [[CrossRef](#)]
40. Kim, H.S.; Seong, J.; Jung, H.J. Optimal domain adaptive object detection with self-training and adversarial-based approach for construction site monitoring. *Autom. Constr.* **2024**, *158*, 105244. [[CrossRef](#)]
41. Rasouli, S.; Alipouri, Y.; Chamanzad, S. Smart Personal Protective Equipment (PPE) for construction safety: A literature review. *Saf. Sci.* **2024**, *170*, 106368. [[CrossRef](#)]
42. Park, S.; Kim, J.; Wang, S.; Kim, J. Effectiveness of Image Augmentation Techniques on Non-Protective Personal Equipment Detection Using YOLOv8. *Appl. Sci.* **2025**, *15*, 2631. [[CrossRef](#)]
43. Jin, X.; Ahn, C.R.; Kim, J.; Park, M. Welding Spark Detection on Construction Sites Using Contour Detection with Automatic Parameter Tuning and Deep-Learning-Based Filters. *Sensors* **2023**, *23*, 6826. [[CrossRef](#)]
44. Hong, Y.J.; Chern, W.C.; Nguyen, T.V.; Cai, H.B.; Kim, H. Semi-supervised domain adaptation for segmentation models on different monitoring settings. *Autom. Constr.* **2023**, *149*, 104773. [[CrossRef](#)]
45. Huang, C.; Zeng, Q.L.; Xiong, F.Y.; Xu, J.Z. Space dynamic target tracking method based on five-frame difference and Deepsort. *Sci. Rep.* **2024**, *14*, 6020. [[CrossRef](#)] [[PubMed](#)]
46. Fang, W.L.; Ma, L.; Love, P.E.D.; Luo, H.B.; Ding, L.Y.; Zhou, A.O. Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology. *Autom. Constr.* **2020**, *119*, 103310. [[CrossRef](#)]
47. Jeon, Y.; Tran, D.Q.; Kulinan, A.S.; Kim, T.; Park, M.; Park, S. Vision-based motion prediction for construction workers safety in real-time multi-camera system. *Adv. Eng. Inform.* **2024**, *62*, 102898. [[CrossRef](#)]
48. Xu, X.B.; Hu, J.C.; Yang, J.; Ran, Y.Y.; Tan, Z.Y. A Fish Detection and Tracking Method Based on Improved Interframe Difference and YOLO-CTS. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–13. [[CrossRef](#)]
49. Zhang, Y.H.; Li, J.C.; Cao, L.C.; Zhou, Q.; Cai, W.; Yu, L.Q.; Li, W.H. Two-staged attention-based identification of the porosity with the composite features of spatters during the laser powder bed fusion. *J. Manuf. Process.* **2024**, *131*, 2310–2322. [[CrossRef](#)]
50. Wang, Z.; Zhu, Y.; Zhang, Y.J.; Liu, S.Y. An effective deep learning approach enabling miners' protective equipment detection and tracking using improved YOLOv7 architecture. *Comput. Electr. Eng.* **2025**, *123*, 110173. [[CrossRef](#)]
51. Xu, K.; Li, F.; Chen, D.J.; Zhu, L.L.; Wang, Q. Fusion of Lightweight Networks and DeepSort for Fatigue Driving Detection Tracking Algorithm. *IEEE Access* **2024**, *12*, 56991–57003. [[CrossRef](#)]

52. Qureshi, A.M.; Alotaibi, M.; Alotaibi, S.R.; AlHammadi, D.A.; Jamal, M.A.; Jalal, A.; Lee, B. Autonomous vehicle surveillance through fuzzy C-means segmentation and DeepSORT on aerial images. *PeerJ Comput. Sci.* **2025**, *11*, e2835. [[CrossRef](#)]
53. Dabash, M.S. Applications of Computer Vision to Improve Construction Site Safety and Monitoring. Ph.D. Thesis, University of Delaware, Newark, Delaware, 2022.
54. Li, Z.-X.; Wang, Y.-L.; Wang, F. DI-YOLOv5: An Improved Dual-Wavelet-Based YOLOv5 for Dense Small Object Detection. *IEEE/CAA J. Autom. Sin.* **2025**, *12*, 457–459. [[CrossRef](#)]
55. Li, F.; Zheng, Y.; Liu, S.; Sun, F.; Bai, H. A multi-objective apple leaf disease detection algorithm based on improved TPH-YOLOV5. *Appl. Fruit Sci.* **2024**, *66*, 399–415. [[CrossRef](#)]
56. Yu, Q.; Han, Y.; Han, Y.; Gao, X.; Zheng, L. Enhancing YOLOv5 Performance for Small-Scale Corrosion Detection in Coastal Environments Using IoU-Based Loss Functions. *J. Mar. Sci. Eng.* **2024**, *12*, 2295. [[CrossRef](#)]
57. Wu, S.; Wang, J.; Wei, W.; Ji, X.; Yang, B.; Chen, D.; Lu, H.; Liu, L. On the Study of Joint YOLOv5-DeepSort Detection and Tracking Algorithm for *Rhynchophorus ferrugineus*. *Insects* **2025**, *16*, 219. [[CrossRef](#)] [[PubMed](#)]
58. Njvisionpower. *Safety-Helmet-Wearing-Dataset (SHWD)*; Github: San Francisco, CA, USA, 2024.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.