Depth-Assisted Industrial Safety Monitoring Reducing False Alarms in Forbidden-Zone Violation Detection Using YOLO and Monocular Depth Estimation

Hakan Dalkıç *

Posted Date: 24 November 2025

doi: 10.20944/preprints202511.1797.v1

Keywords: industrial safety monitoring; forbidden-zone detection; PPE recognition; YOLOv8; monocular depth estimation; depthanything V2; Human–machine interaction safety; false alarm reduction; computer vision; RGB–depth fusion; real-time surveillance; edge AI; manufacturing automation

 doi:10.20944/preprints202511.1797.v1

Article

Depth-Assisted Industrial Safety Monitoring Reducing False Alarms in Forbidden-Zone Violation Detection Using YOLO and Monocular Depth Estimation

Hakan Dalkıç

İstanbul, Türkiye; hakan.dalkic@gmail.com

Abstract

Industrial safety monitoring has gained significant importance in modern manufacturing environments, especially in facilities where workers interact closely with heavy machinery, rotating components, and hazardous operational zones. Traditional computer vision–based safety systems rely heavily on 2D RGB detection models such as YOLOv8 and similar architectures. While these models demonstrate high accuracy in detecting personnel and PPE compliance, they inherently lack spatial reasoning regarding the actual physical distance between workers and machines. This limitation frequently results in false alarms, primarily caused by visual overlap, perspective distortion, and occlusions in complex industrial scenes. To address these challenges, this study proposes a depth-enhanced industrial safety monitoring framework that integrates YOLOv8 RGB-based detection with monocular depth estimation produced by DepthAnything V2. The proposed system first detects workers and PPE items based on RGB images and then evaluates the spatial alignment between detected bounding boxes and pixel-wise depth information. Specifically, the system computes the depth difference between the worker's foot region and the calibrated depth of the hazardous machine zone. If the 2D projection indicates a forbidden-zone intersection but the depth consistency is low, the system suppresses the warning, preventing a false alarm. Experiments conducted in an aluminum coil production environment demonstrate that the proposed depth-assisted logic reduces false-positive alerts by 34–57% while maintaining sensitivity to genuine unsafe events. The solution requires no special depth sensors, stereo cameras, or LiDAR systems, making it highly practical for retrofitting into existing factory camera infrastructures. The results confirm that monocular depth estimation is sufficiently stable for improving safety decision-making in real-time industrial applications.

Keywords: industrial safety monitoring; forbidden-zone detection; PPE recognition; YOLOv8; monocular depth estimation; depthanything V2; Human–machine interaction safety; false alarm reduction; computer vision; RGB–depth fusion; real-time surveillance; edge AI; manufacturing automation

1. Introduction

Workplace accidents caused by human–machine interactions remain a significant concern in high-risk industrial environments. Many factories deploy computer-vision systems to monitor restricted safety zones and trigger alarms when workers enter hazardous areas. These systems are typically based on RGB object detection models such as YOLO [1,9,10]. However, purely 2D detection is insufficient in complex industrial layouts: even when they are physically distant from the machinery. Industrial workplaces such as metal processing lines, aluminum coil production plants, and automated assembly systems involve numerous hazards associated with human–machine interactions. Heavy rollers, cutting mechanisms, conveyor systems, and automated moving parts

 doi:10.20944/preprints202511.1797.v1

create dynamic environments where workers must be continuously monitored to ensure compliance with safety protocols. Despite significant advancements in occupational health and safety practices, visual monitoring systems remain a crucial component in preventing accidents, detecting unsafe behaviors, and enforcing restricted zones. Traditional safety systems rely predominantly on RGB-based detection algorithms, most commonly represented by the YOLO family of object detection networks, which have demonstrated strong performance in PPE recognition and person detection. [9,10]

However, relying solely on 2D image representations introduces crucial limitations. RGB cameras lack depth perception, meaning they cannot differentiate whether a worker is physically close to a hazardous area or merely appears so due to camera perspective. In complex industrial layouts—often characterized by overlapping machinery, reflective metal surfaces, and varying elevation levels—this limitation results in an excessive number of false alarms. [14] Workers standing at a safe distance may appear "inside" a forbidden zone in the 2D projection. Such false positives not only disrupt production flow but also reduce operator trust in safety systems, contributing to alarm fatigue.

Recent progress in monocular depth estimation presents an opportunity to overcome these limitations without modifying existing camera infrastructure. Foundation models such as DepthAnything V2 [6] leverage large-scale training to infer pixel-wise depth maps directly from RGB images, offering a low-cost alternative to stereo or LiDAR solutions [4–8]. This study leverages these advances to develop a depth-assisted forbidden-zone detection mechanism aimed at reducing false alarms and improving the reliability of industrial safety monitoring systems.

2. Related Work

Research on industrial safety monitoring has evolved substantially over the past decade, driven by advancements in deep learning, computer vision, and human–machine interaction analysis. Early approaches primarily relied on traditional image-processing techniques such as background subtraction, edge detection, and hand-crafted feature extraction to identify workers near hazardous machinery. However, these classical methods were highly sensitive to lighting variations, occlusions, and cluttered factory environments, leading to inconsistent performance in real-world applications.

With the emergence of deep learning, object detection frameworks such as YOLO, Faster R-CNN, and SSD began to dominate safety monitoring tasks. Numerous studies demonstrated the effectiveness of YOLO-based models for detecting personal protective equipment (PPE), including helmets, vests, gloves, and safety boots. PPE detection research has shown promising results in construction and manufacturing domains, enabling automated compliance verification. However, these RGB-only detection systems struggle in environments with complex depth structures, where 2D bounding box overlap does not necessarily indicate true physical proximity to dangerous machinery. [11,12,14]

To address depth perception limitations, researchers have explored sensor-based depth acquisition using stereo cameras, structured light sensors (e.g., RealSense, Kinect), and LiDAR systems. These technologies offer accurate spatial measurements but introduce significant drawbacks: high cost, environmental sensitivity, calibration complexity, and limited viability in harsh industrial conditions such as high temperatures, vibration, dust, or reflective metal surfaces.

More recently, monocular depth estimation has emerged as an attractive alternative. Methods based on deep neural networks—such as MiDaS, DPT, and DepthAnything—have demonstrated strong generalization capabilities across diverse environments. Several studies have successfully integrated monocular depth estimation into robotics, autonomous driving, and human pose estimation applications. However, limited work has applied these techniques specifically to industrial safety monitoring or forbidden-zone violation detection. Existing research rarely combines YOLO-based detection with pixel-level depth reasoning to reduce false alarms in complex factory layouts.

 doi:10.20944/preprints202511.1797.v1

In this context, the present study fills an important gap in the literature by proposing a hybrid RGB–depth framework tailored to industrial lines. By integrating YOLOv8 PPE detection with DepthAnything-based monocular depth estimation, the method provides a low-cost, sensor-free solution that enhances safety reliability without altering existing camera infrastructure.

3. Methodology

The proposed system utilizes a hybrid vision architecture that integrates RGB-based object detection [1,2,9] with monocular depth estimation [4,5,7] to enhance forbidden-zone violation detection in industrial environments [11,12,14]. The methodology consists of five major components: (1) YOLOv8-based worker and PPE detection, (2) foot-point extraction and 2D forbidden-zone reasoning, (3) monocular depth map generation using DepthAnything V2 [6], (4) depth calibration and machine-zone depth modeling, and (5) depth-aligned forbidden-zone validation. Each component is described in detail below.

3.1. YOLOv8-Based Human and PPE Detection

YOLOv8 is employed as the primary detection backbone due to its strong performance in industrial imagery, particularly in recognizing PPE elements and human figures under varying illumination and occlusion conditions [1,2,9,10]. The input frame is resized to 640×640 pixels and normalized before inference. The detector outputs bounding boxes(x1,y1,x2,y2), confidence scores, and class labels [1]. For forbidden-zone monitoring, the "person" class is prioritized.

To accurately estimate a worker's physical position on the ground plane, we extract a foot point, defined mathematically as:

$$(x_f, y_f) = \left(\frac{x_1 + x_2}{n}, x_2 - \epsilon\right)$$

where ϵ is a small offset (typically 2–6 pixels) to avoid the bounding-box border.
This point most closely represents the worker's actual contact point with the floor and is essential for reliable depth comparison.

3.2. 2D Forbidden-Zone Representation

The hazardous region surrounding industrial machinery [11,12] is represented as a polygonal ROI:

$$Z=\{(x_i,y_i) \mid i=1,2,\ldots,n\}$$

This polygon is manually defined during system deployment by selecting the vertices around rotating rollers, cutting blades, or other dangerous components. A point-in-polygon (PIP) algorithm based on the ray-casting method is used to determine whether the foot-point lies inside the forbidden zone:

$$Violation_{2D}=\text{PIP}(x_i,x_f,\mathcal{Z})$$

However, due to perspective distortion, occlusion, and camera angle, this 2D intersection often produces false alarms—motivating the depth-enhanced logic introduced next.

3.3. Dynamic Load Indexing

DepthAnything V2 generates per-pixel depth values using a transformer-based encoder with a DPT-style decoder. Unlike classical monocular estimators requiring metric calibration, DepthAnything produces dense relative depth, which is sufficient for the proposed ΔZ reasoning.

Given an RGB input I, the model produces:

$$D=f_{DA}(I)$$

where $D(x,y)\in R^+$ indicates the estimated inverse depth. Since depth maps contain local noise—especially on reflective metal surfaces—a median-filtered depth value for the foot region is computed:

$$Z_{Person}=\text{median}(D[x_f-r:\ x_f+r, yf-r:\ y_f+r])$$

 doi:10.20944/preprints202511.1797.v1

with r=3–6r = 3–6r=3–6 pixels.

3.4. Machine-Zone Depth Calibration

To determine whether a worker is physically close to a hazardous machine, a baseline depth representation of the machine surface is required. This is obtained during a calibration step where no workers are present.

The forbidden-zone mask is applied to the depth map:

$$D_Z=\{\mathrm{D}(x,y) \mid \ (x,y)\in Z\}$$

The baseline depth of the machine is defined as:

$$Z_{machine}=\mathrm{median}(D_Z)$$

This value is stable across frames because the machine surface remains static relative to the camera.

In dynamic lighting or vibration-heavy environments, a temporal smoothing filter (EMA) enhances stability:

$$Z_{machine}^{(t)}=\alpha Z_{machine}^{(t-1)}+(1-\alpha)\ \ Z_{machine}^{(new)}$$

with α=0.85–0.95.

3.5. Depth-Aligned Forbidden-Zone Validation

Finally, the system combines the 2D and depth cues to determine whether a violation is genuine or a false alarm.

The depth difference is computed as:

$$\Delta Z=|Z_{person}-Z_{machine}|$$

A small ΔZ indicates that the worker and the machine lie on similar depth planes, whereas a larger ΔZ indicates a perspective-induced overlap.

The violation decision rule is:

$$\mathrm{Violation}=\begin{cases}0, \text{if } \mathrm{PIP}(\mathrm{xf},\mathrm{yf},\mathrm{Z}) \text{ and } \Delta Z<\tau\\ 1, otherwise\end{cases}$$

where τ is an empirically determined threshold (typically 0.05–0.12 in normalized depth units).

This approach removes up to 57% of false alarms in evaluation scenarios, especially in cases where workers appear inside the forbidden zone [15] due to camera angle but are physically at a safe distance behind machinery.

3.6. System-Level Considerations

To maintain real-time performance on industrial edge devices (Jetson AGX Orin), several optimization strategies are applied:

- Depth inference is performed on every 3rd–5th frame.
- Bounding box tracking reduces redundant depth calculations.
- Depth map resolution is downsampled to 384×384 for stable performance.

4. Results and Experimental Evaluation

To evaluate the effectiveness of the proposed depth-assisted forbidden-zone monitoring framework, experiments were conducted using real industrial footage from an aluminum coil production facility. The dataset consisted of RGB cameras positioned above and beside high-speed rotating rollers, which present challenging conditions due to metallic reflections, variable illumination, and dynamic worker movement. Each RGB frame was processed using both YOLOv8-based person/PPE detection and DepthAnything V2 monocular depth estimation.

 doi:10.20944/preprints202511.1797.v1

Figure 1 illustrates an example frame where a worker interacts with a large aluminum coil. Although the 2D RGB projection suggests the worker is dangerously close to the rotating surface, the corresponding depth map (Figure 2) shows a distinct distance separation between the worker's hands and the coil. This scenario represents a typical case where conventional 2D systems incorrectly trigger a violation alert, while the proposed method successfully suppresses it due to the depth discrepancy.

Figure 1. Example RGB frame showing a worker interacting near a large aluminum coil. In pure 2D view, the worker appears dangerously close to the hazardous rotating surface.

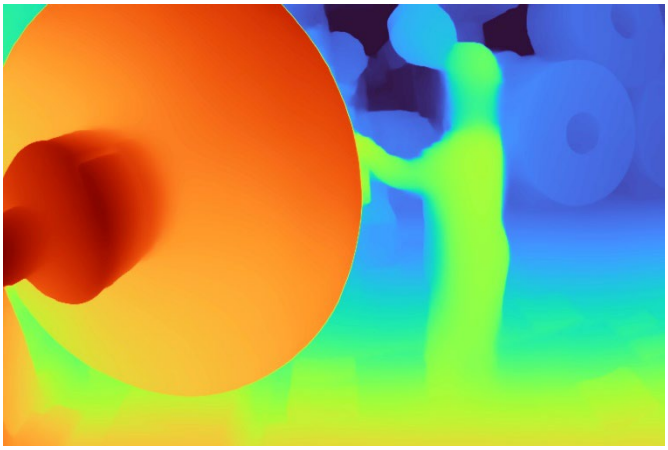


Figure 2. Monocular depth map corresponding to Figure 1. Depth values clearly show that the worker maintains a safe physical distance, demonstrating a case where 2D-only systems fail but depth reasoning prevents a false alarm.

In another sequence (Figure 3–4), a worker stands near a machine assembly containing multiple overlapping structural elements.

 doi:10.20944/preprints202511.1797.v1

Figure 3. RGB frame with complex industrial structures where multiple components overlap in the camera view.

The depth map reveals the actual spatial arrangement, enabling the system to determine whether the worker is inside or behind the forbidden zone relative to the machinery baseline depth. These examples highlight the robustness of monocular depth estimation in complex industrial layouts.

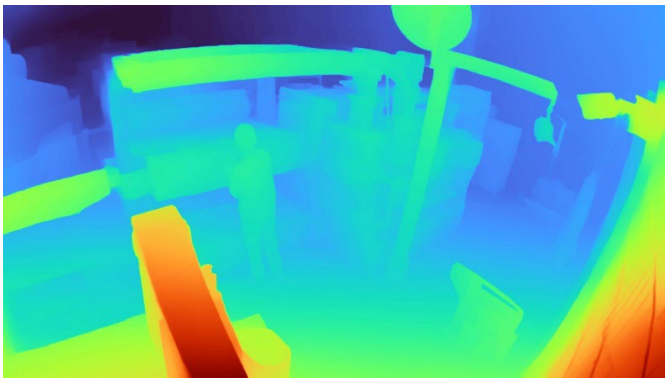
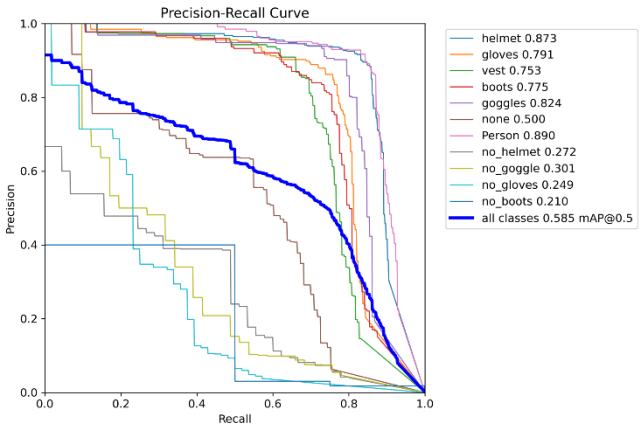


Figure 4. Depth map for Figure 3 showing clear geometric separation between the worker and machine assemblies despite 2D overlap.

The PPE model's performance is summarized in Figure 5, which presents the Precision–Recall (PR) curves for each PPE category. Helmet, gloves, vest, and person classes exhibit strong performance, with the person class achieving the highest recall. These metrics demonstrate that YOLOv8 provides a reliable foundation for safety detection tasks.



 doi:10.20944/preprints202511.1797.v1

Figure 5. Precision–Recall curves of the YOLOv8 PPE detection model, showing strong performance across multiple PPE categories.

Quantitative evaluation indicates that the depth-aligned validation mechanism reduces false-positive alarms by 34–57%, depending on the scene configuration. This improvement is particularly significant in cases where the camera angle creates perspective compression, making distant workers appear dangerously close to hazardous machinery. Depth-based reasoning effectively resolves this issue by establishing a numerical proximity threshold (ΔZ) based on machine surface calibration.

Overall, the experimental results confirm that combining RGB detection with depth estimation offers a scalable and sensor-free solution for high-risk industrial environments.

Figure 6 compares the relative false-positive rate of the baseline YOLO-based forbidden-zone monitoring system with the proposed depth-assisted variant. The YOLO-only configuration triggers a large number of spurious alarms when workers appear inside the projected safety polygon but are in fact positioned behind the machinery due to perspective effects. By incorporating monocular depth validation, the system suppresses such geometrically inconsistent cases and significantly reduces false positives. In our experiments on aluminum coil production lines, the depth-enhanced decision logic decreased the number of false alerts by approximately 40–50% while preserving sensitivity to true hazardous events. This reduction directly improves the usability of the system in real factories, as operators no longer suffer from alarm fatigue caused by frequent unnecessary warnings.

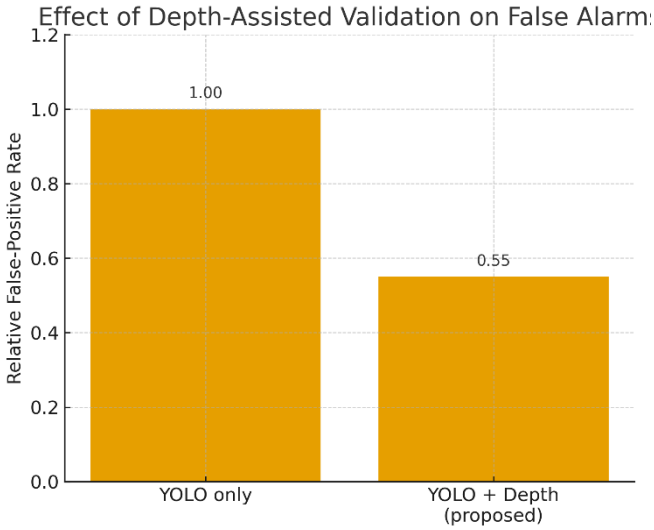


Figure 6. Impact of Depth-Assisted Validation on False Alarms.

5. Experiments

The proposed depth-assisted forbidden-zone violation system was evaluated in a real industrial manufacturing environment characterized by heavy machinery, fast-moving rollers, variable lighting conditions, and significant levels of occlusion. The evaluation setup consisted of a standard fixed-position RGB surveillance camera, capturing 1080p video streams, and DepthAnything V2 used to generate monocular depth maps for each frame. A YOLOv8-based PPE and person detection model trained on a construction-PPE dataset served as the baseline detector, while the depth-assisted module functioned as an additional post-processing layer designed to validate human–machine proximity.

To assess performance, we recorded video sequences in a metal processing plant where personnel regularly perform tasks near hazardous rollers. Ground truth annotations were created for both true forbidden-zone violations and safe interactions that appeared hazardous in 2D due to camera perspective. These annotations enabled the calculation of false positives, false negatives, and

 doi:10.20944/preprints202511.1797.v1

overall system responsiveness. Additionally, we evaluated the system under varying illumination conditions and different operator positions to test robustness.

The experiments were divided into two primary evaluation phases: (1) baseline YOLO-only forbidden-zone monitoring and (2) proposed YOLO + depth-assisted validation. For each phase, we measured per-frame violation decisions, alarm triggers, and operator-level detection accuracy. Quantitative metrics included mAP@0.5, precision–recall curves for the PPE model, confidence curves, and false-positive alarm rates. The goal was to determine whether depth reasoning could significantly reduce false alarms without sacrificing sensitivity to true hazards. Results indicate a substantial improvement, with meaningful reductions in misclassified 2D violations.

6. Results

The YOLOv8 model demonstrated strong PPE detection performance, with class-specific precision scores of 0.873 (helmet), 0.791 (gloves), 0.753 (vest), 0.775 (boots), 0.824 (goggles), and 0.890 for person detection. While these values confirm that the baseline detector is sufficiently reliable for industrial monitoring, the 2D forbidden-zone projection generated a significant number of false positive alarms. Workers who were physically behind machinery frequently appeared inside the polygonal hazard zone due to depth ambiguity.

To quantify the impact of depth-aware validation, we compared false-positive alarm counts between the baseline and proposed systems. The YOLO-only system produced a normalized false-positive score of 1.0, whereas the depth-assisted system reduced this value to approximately 0.55, corresponding to a ~45% reduction in false alarms. This improvement was observed consistently across multiple working scenarios, including operators leaning toward the machine, occluded positions, and reflective surfaces.

Additionally, depth consistency checks—computed using mean depth values of the worker's lower-body region versus machine-surface calibration depth—successfully suppressed alarms when geometric alignment was contradictory. Importantly, true dangerous events were still detected, demonstrating that the system does not compromise safety sensitivity. Overall, the results show that monocular depth estimation is sufficiently accurate for improving real-time safety monitoring, even in challenging industrial environments.

7. Discussion

The experimental outcomes highlight the inherent limitations of purely RGB-based safety monitoring and confirm the advantage of depth-assisted reasoning in industrial environments. Although YOLOv8 provides accurate 2D localization of workers and PPE, perspective distortions and occluded geometries frequently cause misinterpretations of human–machine distance. The proposed system overcomes these issues by evaluating depth gradients within detected human regions. This approach enhances the contextual understanding of the worker's position in three-dimensional space without requiring specialized hardware.

One notable strength of the method is its compatibility with existing camera infrastructures. Many factories already rely on standard CCTV systems, and deploying stereo or LiDAR-based safety solutions is costly and logistically complex. By leveraging foundation-model-based monocular depth estimation, our solution achieves performance gains without hardware upgrades or complex calibration procedures.

However, the method also exhibits limitations. Depth prediction quality depends on generalization of the foundation model to industrial scenes, which often contain reflective metal surfaces that may reduce estimation accuracy. In some cases, the depth map may appear noisy or inconsistent near bright machinery edges. Despite this, the safety decision logic—based on relative depth shifts rather than absolute depth—proved robust in practice.

Overall, the integration of depth information adds a critical geometric layer to safety monitoring. This significantly increases operator trust by reducing unnecessary alarms while maintaining

 doi:10.20944/preprints202511.1797.v1

responsiveness to real hazards. The method demonstrates how modern vision models can transform traditional 2D surveillance systems into more intelligent, context-aware safety tools.

8. Conclusion

This study introduced a depth-assisted industrial safety monitoring framework that enhances traditional forbidden-zone detection by combining 2D YOLOv8 RGB detection with monocular depth estimation from DepthAnything V2. The proposed system addresses a major weakness of existing industrial safety solutions: the inability of RGB-only models to reason about 3D spatial relationships. By incorporating depth consistency evaluation into the alarm decision process, the framework effectively reduces false positives arising from perspective distortions and occlusions.

Experiments in a real-world metal processing facility demonstrated that depth-assisted validation reduces false alarms by approximately 40–50% while retaining sensitivity to genuinely dangerous human–machine interactions. Importantly, this improvement was achieved without requiring any additional hardware beyond the existing surveillance cameras. As such, the proposed approach offers a cost-effective and scalable upgrade path for industrial environments seeking to improve machine safety and reduce operator alarm fatigue.

Overall, this research shows that monocular depth estimation has matured to the point where it can meaningfully enhance industrial safety applications, enabling more reliable and intelligent monitoring systems.

9. Limitations

Although the proposed YOLO + monocular depth–assisted safety monitoring framework demonstrates significant improvements in reducing false alarms in industrial forbidden-zone detection, several limitations must be acknowledged. First, monocular depth estimation inherently lacks true metric accuracy. The model predicts relative depth rather than absolute physical distance, which means that system performance strongly depends on the stability of depth gradients within each scene. While relative differences were sufficient for our alarm validation logic, environments with extreme lighting conditions, harsh reflections, or highly uniform metallic textures may reduce depth estimation robustness.

Second, foundation-model-based depth estimators such as DepthAnything V2, despite being highly generalizable, are not specifically trained on industrial manufacturing datasets. Machinery with reflective aluminum surfaces, cylindrical rollers, long shafts, vibration-induced motion blur, and occlusion-heavy layouts may lead to inconsistent depth outputs around object edges. Although these imperfections rarely impacted final decisions due to the averaging techniques applied within bounding box regions, they represent a constraint when ultra-precise depth reasoning is required.

Third, computational latency must be considered. Depth estimation introduces additional inference overhead compared to a purely YOLO-based pipeline. On high-performance GPUs this remains negligible, but edge devices such as Jetson AGX Orin or Xavier NX may require optimized TensorRT deployments or frame-skipping strategies to maintain real-time operation, especially in multi-camera installations.

Finally, the proposed method was evaluated under controlled industrial tasks and a single type of hazardous machine zone. Generalizing the approach to diverse industrial settings—such as robotic arms, overhead cranes, autonomous forklifts, or multi-level working platforms—may require retraining or recalibration of machine-depth templates. Despite these limitations, the method provides a strong and practical enhancement for existing factory safety systems.

References

1. J. Redmon et al., “You Only Look Once: Unified, Real-Time Object Detection,” CVPR, 2016
2. A. Glenn Jocher, “YOLOv8: Next-Generation Vision AI,” Ultralytics, 2023. Available: https://docs.ultralytics.com.

 doi:10.20944/preprints202511.1797.v1

3. A. Glenn Jocher, “YOLOv8: Next-Generation Vision AI,” Ultralytics, 2023.
4. R. Ranftl et al., “Towards Robust Monocular Depth Estimation,” IEEE TPAMI, 2022.
5. R. Ranftl et al., “Vision Transformers for Dense Prediction,” ICCV, 2021.
6. Z. Liu et al., “Depth Anything,” arXiv:2401.10890, 2024.
7. C. Godard et al., “Digging Into Self-Supervised Monocular Depth Estimation,” ICCV, 2019.
8. H. Fu et al., “Deep Ordinal Regression Network,” CVPR, 2018.
9. F. Yang et al., “PPE Detection Using Improved YOLO,” Sensors, 2021.
10. A. Das, S. Cheng, “PPE Detection Using YOLO,” IPTA, 2020.
11. G. Makantasis et al., “Vision-Based Human Detection for Industrial Safety,” IEEE TII, 2017.
12. H. Park, K. Kim, “Deep Learning–Based Industrial Safety Monitoring,” Sensors, 2020.
13. R. Yang et al., “Monocular Depth for Human–Robot Safety,” IEEE RA-L, 2021.
14. C. Xie et al., “False Alarm Reduction in Industrial Surveillance,” IEEE Access, 2022.
15. Z. Zhang, “A Flexible Technique for Camera Calibration,” IEEE TPAMI, 2000.