



OPEN YOLOv10-based multi-scale variant object detection for multi-category PPE non-compliance monitoring on construction sites

Seunghyeon Wang¹, Hyuna Kim², Jiyoung Yeo³ & Juhung Kim⁴✉

Multi-category Personal Protective Equipment (PPE) non-compliance monitoring on construction sites remains challenging because visual cues for missing PPE often appear at multiple scales, particularly as small objects. To address these challenges, an enhanced YOLOv10-based detector is developed that incorporates Omni-Dimensional Dynamic Convolution (ODConv) to strengthen fine-grained feature representation, an Efficient Multi-scale Attention-guided Bidirectional Feature Pyramid Network (EMA-BiFPN) to improve cross-scale fusion, a four-scale prediction strategy to increase sensitivity to extremely small instances, and the Minimum Points Distance IoU (MPDIoU) loss to stabilize bounding-box regression. Experiments are conducted on a dataset of 4060 images annotated for five violation categories (No-helmet, No-mask, No-vest, No-glove, and No-shoes), using stratified splits and a controlled hyperparameter search (243 configurations). On the test set, the proposed method achieves 92.49 mAP@50 and 70.04 mAP@50:95 while maintaining 57.28 FPS, outperforming representative two-stage, one-stage, and transformer-based baselines. The largest gains are observed in small-object regimes, reaching $AP_{ES} = 82.74$ and $AP_S = 88.96$, indicating improved robustness to distant and tiny PPE cues. Robustness analyses further show that performance is less sensitive to weight initialization than to dataset partitioning, and condition-based evaluation confirms that far-field views and worker overlap remain the dominant failure cases due to the difficulty of localizing extremely small instances.

Keywords Personal protective equipment monitoring, Small object detection, Image processing, Deep learning, You only look once version 10

Personal Protective Equipment (PPE)—such as safety helmets, high-visibility vests, protective gloves, face masks, and safety shoes—plays a fundamental role in preventing injuries on construction sites. However, workers do not always comply with PPE requirements, often due to discomfort, time pressure, or limited safety awareness¹. Maintaining consistent compliance among all personnel is therefore essential for improving safety performance. In practice, PPE enforcement often relies on manual inspections, routine patrols, or lengthy reviews of surveillance recordings. These approaches are labor-intensive, difficult to scale, and susceptible to human error, particularly when prolonged monitoring leads to inspector fatigue².

Deep-learning-based computer vision offers a practical route for automating this process. Modern object detectors can identify and localize predefined targets using bounding boxes, enabling automated screening of compliance indicators in images collected from fixed cameras or wearable devices. When trained on datasets covering diverse viewpoints, lighting conditions, worker postures, and object sizes, such models can achieve more robust performance across real-world scenes³. Within this context, the You Only Look Once (YOLO) family has been widely adopted for construction safety applications because of its real-time inference capability and continued architectural improvements across successive versions^{4–6}. Accordingly, a growing body of research has introduced YOLO-based variants for safety monitoring, showing that detector-level refinements can yield measurable performance gains.

Nevertheless, multi-category PPE non-compliance detection remains challenging in operational environments. Evidence of PPE absence is not concentrated in a single body region but is distributed across

¹Institute for Environmental Design and Engineering, University College London, London WC1H 0NN, UK.

²Department of Pediatrics, College of Medicine, Hanyang University, Seungdong-gu, Seoul 133791, Korea.

³Department of Nursing, College of Nursing, Hanyang University, Seungdong-gu, Seoul 133791, Korea.

⁴Department of Architectural Engineering, Hanyang University, Seungdong-gu, Seoul 133791, Korea. ✉email: kcr97jhk@hanyang.ac.kr

multiple locations and often appears at very small scales. For example, violations related to helmets and masks depend on cues around the head and face, whereas missing gloves and safety shoes must be inferred from the hands and feet—regions that are frequently obscured by tools, materials, body pose, or other workers. In wide-field or distant imagery, these regions may occupy only a few pixels, making them highly vulnerable to feature loss during repeated downsampling. These factors reduce feature quality at small scales and can lead to unstable localization under occlusion, crowding, and dense worker interactions^{7–9}. Despite these practical difficulties, detection architectures explicitly tailored to small, occluded, and multi-scale PPE violations remain relatively limited.

Although modules such as Omni-Dimensional Dynamic Convolution (ODConv), BiFPN-style feature fusion, and IoU-based regression refinements have been explored in prior object-detection studies, this study does not claim novelty from any single standalone component. Rather, the contribution lies in a task-specific adaptation of YOLOv10 for multi-category PPE non-compliance detection on construction sites. Compared with prior ODConv- or BiFPN-based detectors developed for other domains, the proposed framework is specifically configured to address four task-driven challenges in PPE monitoring: (1) weak and fine-grained visual cues from localized body regions, (2) loss of small-object information during feature extraction and multi-scale fusion, (3) unstable localization of tiny or overlapping violation instances, and (4) insufficient prediction resolution for extremely small distant workers.

To address these limitations, this study develops an enhanced YOLOv10-based framework for automated PPE non-compliance monitoring across five violation categories: no-helmet, no-mask, no-vest, no-glove, and no-shoes. First, ODConv is incorporated into the backbone to enable content-adaptive feature extraction across spatial positions, channels, and kernel settings, thereby improving responsiveness to small and partially visible cues under clutter and scale variation. Second, an Efficient Multi-scale Attention-guided Bidirectional Feature Pyramid Network (EMA-BiFPN) is introduced to strengthen cross-scale information exchange, allowing shallow spatial details and deeper semantic features to be fused more effectively. Third, the Minimum Points Distance Intersection over Union (MPDIoU) loss is adopted to provide stronger geometric constraints for bounding-box regression, thereby improving localization stability for small and highly overlapping instances. Finally, a four-scale prediction design is employed to increase sensitivity across extremely small, small, medium, and large targets, reflecting the wide scale variation of PPE-related cues in real construction imagery.

Accordingly, the novelty of this study lies not in proposing an entirely new detection algorithm, but in the problem-driven integration and evaluation of complementary architectural modifications within YOLOv10 for PPE non-compliance monitoring. The main contributions are summarized as follows:

- (1) A task-specific YOLOv10 adaptation is presented for multi-class PPE non-compliance detection by integrating ODConv, EMA-BiFPN, MPDIoU, and a four-scale prediction strategy.
- (2) Hyperparameter effects are systematically examined, and robustness to stochastic influences—such as random initialization and data-splitting variation—is quantified across violation categories.
- (3) Generalization from validation to test sets is assessed, sensitivity to imaging conditions is analyzed, and an ablation study is conducted.
- (4) The proposed approach is benchmarked against representative object detectors to demonstrate its competitive accuracy and efficiency.

Related work

Vision-based studies on non-PPE monitoring in construction

A substantial body of work has explored vision-based PPE/non-PPE monitoring on construction sites using deep-learning detectors, most commonly CNN-based object detection frameworks. Earlier studies frequently adopted two-stage architectures. For example, Fang et al.¹⁰ applied Faster R-CNN and achieved high performance (90.1–98.4%) under multiple scenarios, but the reported inference time was about 0.2 s per image, which can hinder real-time operation. Gu et al.¹¹ further improved Faster R-CNN by incorporating multi-scale training and expanding anchor settings, reporting an additional accuracy gain of roughly 7%, although runtime efficiency was not detailed.

Because practical monitoring systems must balance detection accuracy with computational cost, many subsequent studies shifted toward one-stage detectors. Yang et al.¹² enhanced YOLOv3 (with a DarkNet53 backbone) by introducing an image-pyramid strategy to better retain multi-scale information, achieving 95.13% accuracy with 0.017 s latency per image and outperforming SSD and Faster R-CNN. Yan and Wang¹³ modified the YOLOv3 backbone using dense connections, improving processing speed by 52% and increasing accuracy by 5.7% while preserving efficient inference. Shen et al.¹⁴ proposed a hybrid pipeline in which faces were first detected via VGG16, after which helmet status was determined using regression-based region mapping and DenseNet classification. Nath¹⁵ initially detected workers using YOLOv3 and then classified PPE using VGG16, ResNet, and Xception, reporting 71.23% accuracy.

More recent work has investigated stronger YOLO variants and broader PPE categories. Wang et al.¹⁶ compared YOLOv5 models, showing that YOLOv5x achieved higher accuracy (86.55%), whereas YOLOv5s enabled faster inference (0.019 s per frame). Lee et al.¹⁷ developed a real-time approach based on YOLACT with a MobileNetV3 backbone, reporting 66.4% accuracy for helmet and vest detection. Nguyen et al.¹⁸ optimized YOLOv5 using the Seahorse Optimization algorithm (SHO), obtaining an average accuracy of approximately 66% across multiple PPE-related categories (helmets, masks, gloves, vests, and shoes). Park et al.¹⁹ analyzed individual augmentation strategies and combined YOLOv8 with transformer backbones to improve non-PPE detection. More recently, Wang et al.²⁰ introduced YOLOv10-based helmet/no-helmet detectors enhanced with transformer-based variants. Related work by Wang²¹ examined non-PPE detection in surveillance and body-

worn imagery using benchmark datasets, while Wang²² proposed a domain-adaptive Faster R-CNN framework to mitigate class imbalance, with particular attention to underrepresented non-helmet classes.

Overall, prior research indicates that deep-learning detectors can perform well within their respective experimental settings. However, in operational construction footage, PPE-related cues often appear at very small scales, and performance can degrade due to strong scale variation, repeated downsampling, dense crowds, and frequent occlusion. Consequently, approaches that explicitly target small-object behavior—together with evaluations that report size-aware metrics—remain comparatively limited in the broader literature on construction PPE non-compliance monitoring.

YOLO-oriented research for small-object detection

There is no single universally accepted definition of a “small object,” because apparent object size depends on image resolution, camera distance, viewing angle, and the evaluation protocol. Many benchmarks therefore use a practical pixel-based convention; for instance, objects with bounding-box areas below 32×32 pixels are commonly treated as small instances²³.

Within the YOLO family, numerous variants have been proposed, and attempts to improve small-object performance in YOLO-style detectors can be summarized into three major directions. First, many studies emphasize preserving spatial detail by reinforcing high-resolution feature pathways and strengthening multi-scale fusion through pyramid-style neck designs, such as Feature Pyramid Network (FPN), Path Aggregation Network (PANet), and Bidirectional Feature Pyramid Network (BiFPN). As one example, Chao et al.²⁴ proposed a YOLOv8 variant incorporating a Deep Spatial Pyramid Fusion (DSPF) module that reduces redundancy and computation associated with residual structures while maintaining effective multi-scale aggregation in lightweight settings.

Second, other methods aim to amplify weak representations of tiny targets by adding lightweight attention modules—such as Squeeze-and-Excitation (SE) and the Convolutional Block Attention Module (CBAM)—or by adopting dynamic convolution strategies such as ODConv. In this line of work, He and Wei²⁵ improved wear-particle recognition by embedding CBAM into a YOLOv5-based detector to emphasize informative features.

Third, localization under stricter IoU conditions has been improved using IoU-aware regression objectives. Su et al.²⁶ compared several IoU-based losses and found that Generalized IoU (GIoU) can outperform the standard IoU loss in localization quality, although the degree of improvement can vary across datasets (e.g., PASCAL VOC). More recently, Zou and Liu²⁷ combined YOLOv10 with ODConv and an MPDIoU-based regression loss, reporting improved precision and recall for traffic-sign detection involving irregular shapes.

Despite these developments, small-object detection remains difficult. Tiny instances often yield weak activations after repeated downsampling, may lose discriminative detail during multi-scale fusion, and can be removed by post-processing in crowded scenes. These issues are particularly pronounced for worker-centered PPE non-compliance monitoring, where evidence of violations is distributed across multiple body regions (e.g., head/face for helmets and masks, torso for vests, and hands/feet for gloves and safety shoes) and is frequently partially visible due to occlusion and overlap. In wide-angle surveillance imagery, such cues can occupy only a few pixels, amplifying the impact of scale variation and clutter.

Proposed approach

Overview of the YOLOv10 baseline for multi-category non-PPE detection

YOLOv10 is a recent one-stage detector designed for end-to-end object detection without relying on Non-Maximum Suppression (NMS). This NMS-free inference is enabled by a consistent dual-assignment strategy that aligns training and inference objectives²¹. As illustrated in Fig. 1, YOLOv10 follows the standard backbone-neck-head design.

The backbone constructs a hierarchical feature representation through lightweight convolutional building blocks (e.g., C2f, SCDowN, and C2fCIB) together with PSA-based attention, producing multi-level feature maps with increasing semantic abstraction (Fig. 1). Shallow stages retain fine spatial detail that is important for small-instance localization, whereas deeper stages emphasize higher-level semantics. The neck then aggregates these multi-resolution features through an FPN-PAN-style pathway, combining high-resolution spatial cues with deeper semantic information.

The detection head adopts a lightweight decoupled structure for classification and box regression and leverages the dual-assignment strategy to support end-to-end prediction. In the baseline setting, YOLOv10 is configured for multi-class non-PPE detection by predicting the following violation categories: no-helmet, no-mask, no-vest, no-glove, and no-shoes. This formulation directly supports compliance assessment because each predicted box corresponds to a specific type of PPE absence observable in the scene.

Architectural enhancements of YOLOv10 for non-PPE compliance monitoring

In real construction imagery, non-PPE cues are often extremely small—particularly in wide-angle surveillance views or long-range shots—because relevant evidence may be confined to localized body regions (e.g., head/face for helmets and masks, torso for vests, and hands/feet for gloves and safety shoes). Under these conditions, the target region may occupy only a few pixels and is easily degraded by (i) repeated downsampling during feature extraction, (ii) smoothing and information dilution during multi-scale fusion, and (iii) suppression effects during inference, especially in crowded scenes with heavy overlap. These difficulties are further amplified by frequent occlusions caused by tools, scaffolding, materials, or inter-person interactions, which can lead to missed detections, inaccurate localization, and confusion between visually similar patterns.

To address these limitations, an enhanced YOLOv10 variant is developed (Fig. 2), with the key modifications summarized in Fig. 3. Four complementary upgrades are introduced:

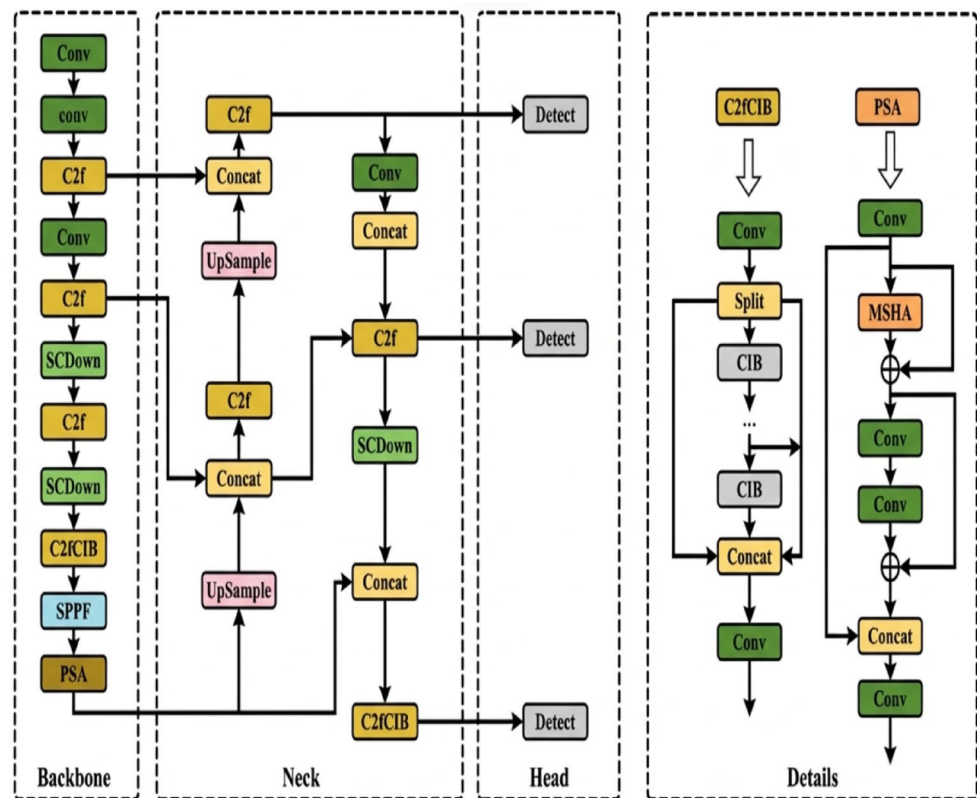


Fig. 1. Baseline YOLOv10 network architecture.

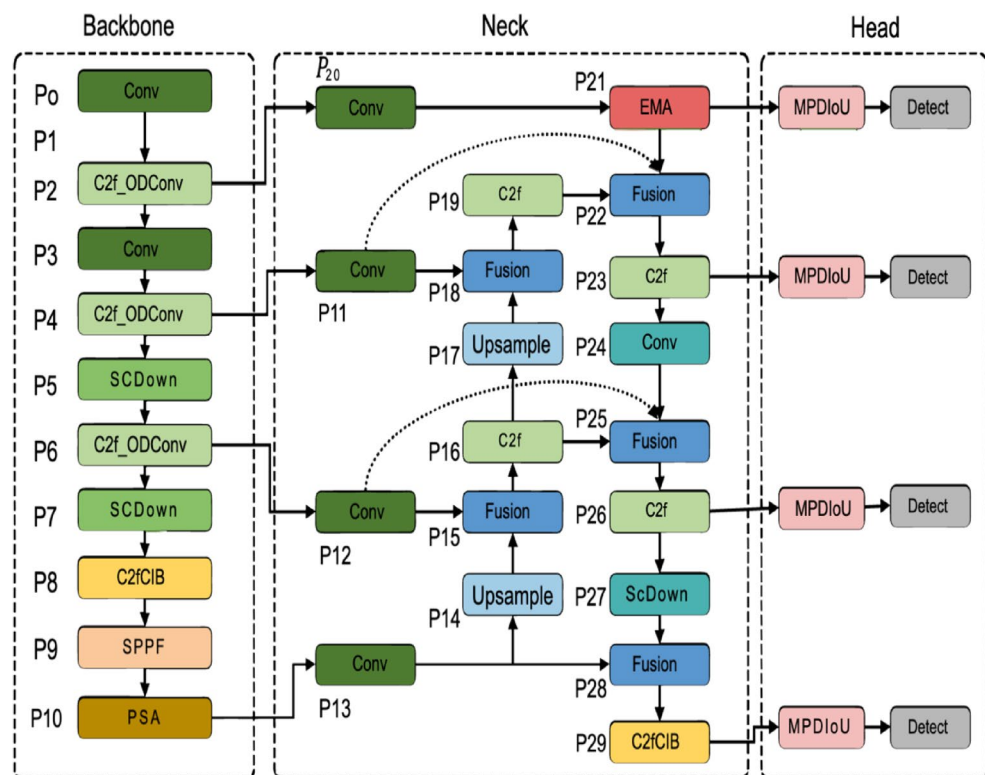


Fig. 2. Overview of the proposed YOLOv10 adaptations for non-PPE detection.

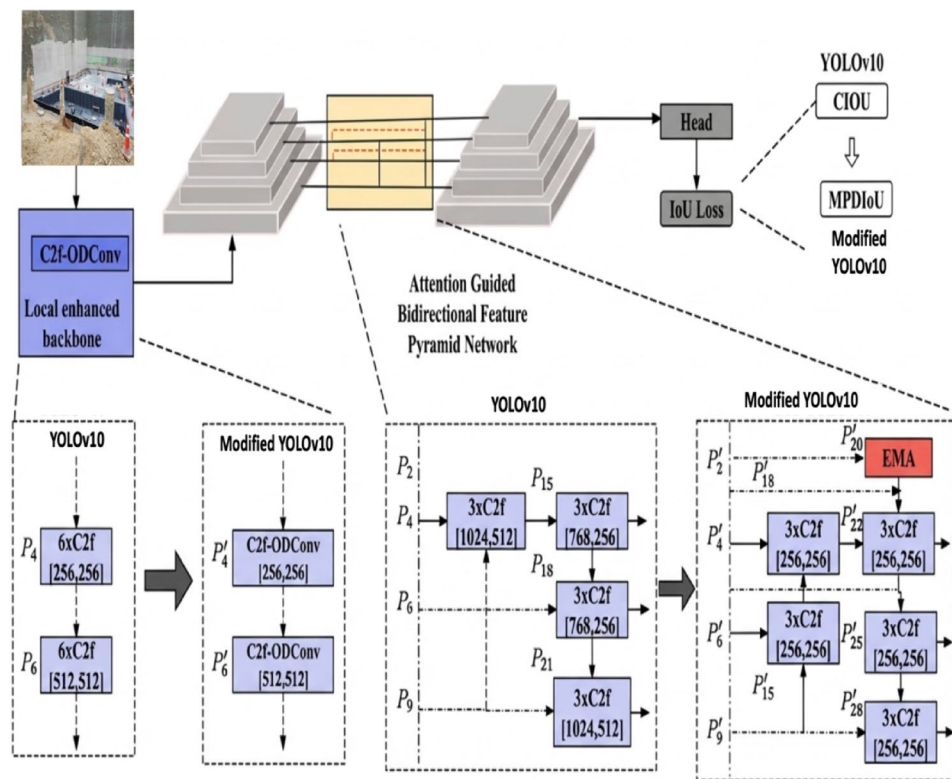


Fig. 3. Side-by-side comparison of the baseline and modified YOLOv10 architectures.

- (1) ODConv in the backbone to strengthen content-adaptive feature extraction for small, cluttered non-PPE cues under scale variation.
- (2) EMA-BiFPN neck to improve bidirectional multi-scale fusion while emphasizing weak, fine-grained features relevant to non-PPE evidence.
- (3) MPDIoU regression loss to stabilize localization learning for small bounding boxes where minor coordinate deviations can cause large IoU changes.
- (4) A four-scale prediction-head design, adding a higher-resolution head to better detect extremely small instances.

Together, these changes target major failure modes in non-PPE monitoring: loss of fine detail after downsampling, weakened propagation of small-scale cues across the pyramid, and unstable regression signals for tiny or heavily overlapping boxes.

ODConv-enhanced backbone for fine-grained non-PPE cues

Visual evidence of PPE non-compliance is often weak, fragmented, and easily obscured. For example, the absence of a helmet or mask may only be reflected through faint contour information, exposed hair or skin regions, or incomplete boundary patterns. Likewise, missing gloves or safety shoes may be only partly visible because of body pose, self-occlusion, or surrounding objects. These weak signals can be confused with background structures such as concrete textures, reflective surfaces, clothing folds, or scaffold edges. Since these cues are also frequently very small, even minor changes in viewpoint or illumination can noticeably alter their appearance.

To improve feature extraction under these challenging conditions, ODConv is incorporated into the C2f block. Unlike standard convolution, ODConv dynamically adjusts the convolution process across four dimensions: spatial position, input channel, output channel, and kernel index. This allows the effective convolutional response to change adaptively according to the input content. Compared with earlier dynamic convolution approaches such as CondConv and DyConv, which usually modulate only part of the convolution operation, ODConv provides a richer form of adaptive kernel selection while still remaining efficient enough for real-time detection^{28,29}. This is especially suitable for non-PPE monitoring, where the target cues are often weak, localized, and highly dependent on scene context.

The ODConv operation can be expressed as:

$$y = (a_{w1}W_1 + \dots + a_{wn}W_n) * x \quad (1)$$

where x and y denote the input and output feature maps, respectively; W_n is the n -th convolution kernel; and a_{wn} is its corresponding weight. With omni-dimensional attention, each kernel is further modulated by multiple attention components:

$$y = (a_{w1} \odot a_{f1} \odot a_{c1} \odot a_{s1} \odot W_1 + \dots + a_{wn} \odot a_{fn} \odot a_{cn} \odot a_{sn} \odot W_n) * x \quad (2)$$

where $\odot$ denotes element-wise multiplication, and a_{fn} , a_{cn} , and a_{sn} correspond to attention terms related to feature-response adaptation, channel-wise emphasis, and spatial weighting, respectively. Through this multi-dimensional modulation, ODConv can amplify informative non-PPE cues while suppressing distracting background patterns.

EMA-BiFPN for attention-guided, bidirectional multi-scale fusion

Non-PPE indicators appear across a broad range of scales because worker distance, camera location, and body posture vary widely in construction imagery. Very small targets, such as distant facial, hand, or foot regions, require high-resolution feature maps to preserve spatial detail, yet such maps often lack strong semantic information. In contrast, deeper layers provide richer semantic context but are spatially coarse and may fail to retain fine details after repeated downsampling. As a result, effective non-PPE detection depends on a fusion mechanism that can inject semantic information into detailed high-resolution features without overwhelming weak small-scale signals.

To address this issue, an EMA-BiFPN structure is introduced by combining bidirectional feature-pyramid fusion with EMA-based attention refinement²⁹. The EMA module strengthens feature representation by jointly capturing channel relationships and spatial dependencies, enabling the network to emphasize regions that are more relevant to PPE-related cues during fusion. At the same time, the BiFPN-style design promotes repeated top-down and bottom-up information exchange, helping spatially detailed shallow features and semantically rich deep features interact more effectively. Shortcut connections are also included between intermediate fusion nodes to reduce information loss and better preserve small-scale features throughout the pyramid³⁰.

Given an input feature map $X \in \mathbb{R}^{H \times W \times C}$, EMA partitions channels into G groups X_g for efficiency. Two lightweight branches (e.g., 1×1 and 3×3 convolutions) capture complementary context. Channel attention is derived via global average pooling:

$$G_{avg}^{1 \times 1} = GAP(X_g^{1 \times 1}), G_{avg}^{3 \times 3} = GAP(X_g^{3 \times 3}) \quad (3)$$

and group-wise attention refinement is computed as:

$$X_g^{EMA} = G_{avg}^{1 \times 1} \odot X_g^{1 \times 1} + G_{avg}^{3 \times 3} \odot X_g^{3 \times 3} \quad (4)$$

Concatenating all X_g^{EMA} along the channel dimension yields the final refined representation. This attention-guided fusion is designed to improve cross-scale context while maintaining computational efficiency suitable for real-time monitoring.

MPDIoU-based regression to stabilize localization for tiny instances

Accurate localization is critical for non-PPE categories because small coordinate offsets can substantially reduce IoU and may also affect classification confidence due to misalignment between the predicted region and the true visual evidence. For tiny targets, overlap-based losses can become unstable: even a 1–2 pixel shift may cause a large overlap drop, weakening the regression signal needed for precise box refinement.

YOLOv10 commonly uses Complete IoU (CIoU)³¹:

$$L_{CIOU} = 1 - IoU(B_{gt}, B_{prd}) + \frac{\rho^2(c_{gt}, c_{prd})}{c^2} + \alpha v \quad (5)$$

where $IoU(B_{gt}, B_{prd}) = \frac{B_{gt} \cap B_{prd}}{B_{gt} \cup B_{prd}}$, $\rho(c_{gt}, c_{prd})$ is the Euclidean distance between box centers, and c is the diagonal length of the smallest enclosing box. The aspect-ratio term is defined as:

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w_{gt}}{h_{gt}} - \arctan \frac{w_{prd}}{h_{prd}} \right)^2, + \alpha = \frac{v}{(1 - IoU(B_{gt}, B_{prd})) + v} \quad (6)$$

Although CIoU introduces geometric constraints beyond overlap, its signal can still be weak when overlap is ambiguous for very small boxes. To strengthen regression for small non-PPE instances, MPDIoU is adopted³². MPDIoU penalizes distances between corresponding corner points, providing additional supervision beyond IoU. For two boxes A and B, the squared distances between matched corners are:

$$d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2 \quad (7)$$

$$d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2 \quad (8)$$

and MPDIoU is computed as:

$$MPDIoU = \frac{A \cap B}{A \cup B} - \frac{w^2 + h^2}{d_1^2 + w^2 + h^2} - \frac{w^2 + h^2}{d_2^2 + w^2 + h^2} \quad (9)$$

where w and h denote the width and height of the box in image coordinates. By incorporating corner-distance constraints, MPDIoU can provide a stronger and more stable learning signal for tiny boxes, improving both position and scale estimation under occlusion and overlap.

Four-scale prediction heads for extremely small non-PPE instances

In compliance monitoring, the hardest cases are often the most safety-critical: distant workers may appear small in surveillance footage, and cues for masks, gloves, and safety shoes can be especially tiny and frequently occluded. If predictions are generated only from coarser feature maps, these cues may collapse into a few activations, increasing the likelihood of missed detections or merged boxes in dense scenes.

To increase sensitivity at the finest scale, the detector is extended to a four-head multi-scale prediction scheme. An additional high-resolution pyramid level is derived from shallower neck features that preserve richer spatial detail, refined via bidirectional fusion, and connected to an extra detection head. As a result, predictions are produced at four resolutions intended to cover extremely small, small, medium, and large instances, respectively. During training, losses are aggregated across all heads:

$$L_{Total} = \sum_{s=1}^4 \left(L_{Cls}^{(s)} + L_{Obj}^{(s)} + \lambda L_{Reg}^{(s)} \right) \quad (10)$$

where $s \in \{1, 2, 3, 4\}$ indexes the prediction head from the highest resolution (targeting extremely small instances) to the lowest resolution (targeting larger instances), $L_{Cls}^{(s)}$ and $L_{Obj}^{(s)}$ are the classification and objectness losses, and λ balances the regression term. Regression is defined using MPDIoU:

$$L_{Reg}^{(s)} = 1 - MPDIoU \left(B_{gt}^{(s)}, B_{prd}^{(s)} \right) \quad (11)$$

where $B_{gt}^{(s)}$ and $B_{prd}^{(s)}$ represent the matched ground-truth and predicted boxes assigned to head s . The added high-resolution head, together with MPDIoU-based supervision, is intended to reduce missed detections for extremely small non-PPE cues while maintaining robustness across scales.

Evaluation metrics

AP

Average Precision (AP) is a standard metric for object detection that summarizes the precision–recall relationship by integrating precision over the recall interval $[0, 1]$ under a given IoU threshold. In this study, AP is computed as:

$$AP = \sum_{i=1}^n p(i) \Delta r(i) \quad (12)$$

where detections are ranked by confidence, $p(i)$ is the precision after the top i predictions are considered, and $\Delta r(i)$ is the recall increment from rank $i-1$ to i . AP50 denotes AP at IoU = 0.50, which allows relatively tolerant matching. AP50:95 is more stringent and is obtained by averaging AP over IoU thresholds from 0.50 to 0.95 in steps of 0.05. Because the latter rewards tighter localization—particularly important in crowded scenes with overlapping instances—both AP50 and AP50:95 are reported.

Size-specific average precision

To analyze performance across different object scales, size-dependent AP is additionally reported. All images are resized to 640×640 using aspect-ratio-preserving letterbox resizing, and bounding-box sizes are measured in this resized coordinate system. For a ground-truth box with width w and height h (in pixels), the size indicator is defined as:

$$s = \min(w, h) \quad (13)$$

Using s , instances are divided into four groups: Extremely Small (ES) if $s < 8$, Small (S) if $8 \leq s < 16$, Medium (M) if $16 \leq s < 32$, and Large (L) if $s \geq 32$. For each subset, AP_{ES} , AP_S , AP_M , and AP_L are computed by restricting evaluation to instances within the corresponding size range, following the same AP procedure used for the primary results (AP50) in this work.

Detection speed

Inference efficiency is measured as the processing time per image and is reported as Frames Per Second (FPS), i.e., the number of images processed per second. This metric reflects computational practicality and indicates whether a model is suitable for real-time construction-site PPE non-compliance monitoring.

Experiments

Data preparation

Dataset sources and original data-collection protocols

The data used in this study were obtained from a previously released body-worn camera dataset and three publicly available datasets. The first source was the body-worn camera dataset, which was originally collected using a

GoPro HERO9 Black camera mounted on the safety helmet of construction workers across five construction sites, as described in our previous publications^{20,21}. In the original data-collection protocol, video was recorded continuously, and frames were sampled at fixed intervals to capture typical construction-site working conditions. Fixed surveillance-camera footage was not used for the multi-category non-PPE experiments because, at common viewing distances, it typically provides reliable cues only for helmet status, whereas other PPE items, such as masks, vests, gloves, and safety shoes, are often indistinguishable. The original on-site data collection was conducted with permission from the participating construction companies, and informed consent was obtained from the participating workers, as reported in the previous publications^{20,21}.

In the present study, no new image or video data were collected, and there was no direct interaction with human participants. All methods in the present study, including dataset selection, curation, annotation processing, model training, and evaluation, were performed in accordance with the relevant guidelines and regulations. All previously released and publicly available datasets were used in accordance with the original sources' stated conditions and applicable licenses/terms of use.

Informed consent was obtained from all identifiable participants, or their legal guardians where applicable, for the publication of the relevant images and related information in an online open-access publication. No participant names, patient names, or direct personal identifiers are included in the manuscript, figures, tables, or supplementary materials.

The sampling procedure initially yielded 53,875 frames, from which 3197 representative images were curated to support a balanced distribution of PPE-compliant and non-compliant cases. To broaden scene diversity and contextual variability, three public datasets were additionally incorporated: the Safety Helmet Wearing Dataset (SHWD)³³, the Safety Helmet Computer Vision Project (SHCVP)³⁴, and the Hard Hat Workers Dataset (HHWD)³⁵.

Representative samples from the four dataset sources are shown in Fig. 4. Because these public datasets primarily provide helmet-related annotations, images in which other PPE items were not visible were excluded from the multi-category non-PPE analysis. Table 1 summarizes the data sources and image counts used to form the experimental dataset. In total, the final dataset comprises 4060 images.

Annotation

For annotation, each occurrence of missing PPE was labeled using rectangular bounding boxes for five categories: no-helmet, no-mask, no-glove, no-vest, and no-shoes. All image regions not associated with these categories were treated as background. Because the detector was designed to localize only PPE violation categories, compliant PPE use was not annotated as a separate class and was instead treated as background during training



Fig. 4. Representative samples from each dataset source.

Dataset name	Number of images	Source
Body-worn camera	1172	Construction sites
SHWD	632	Construction sites + general image
SHCVP	1121	Construction sites
HHWD	1135	Construction sites + general image

Table 1. Dataset overview and descriptive statistics by source.

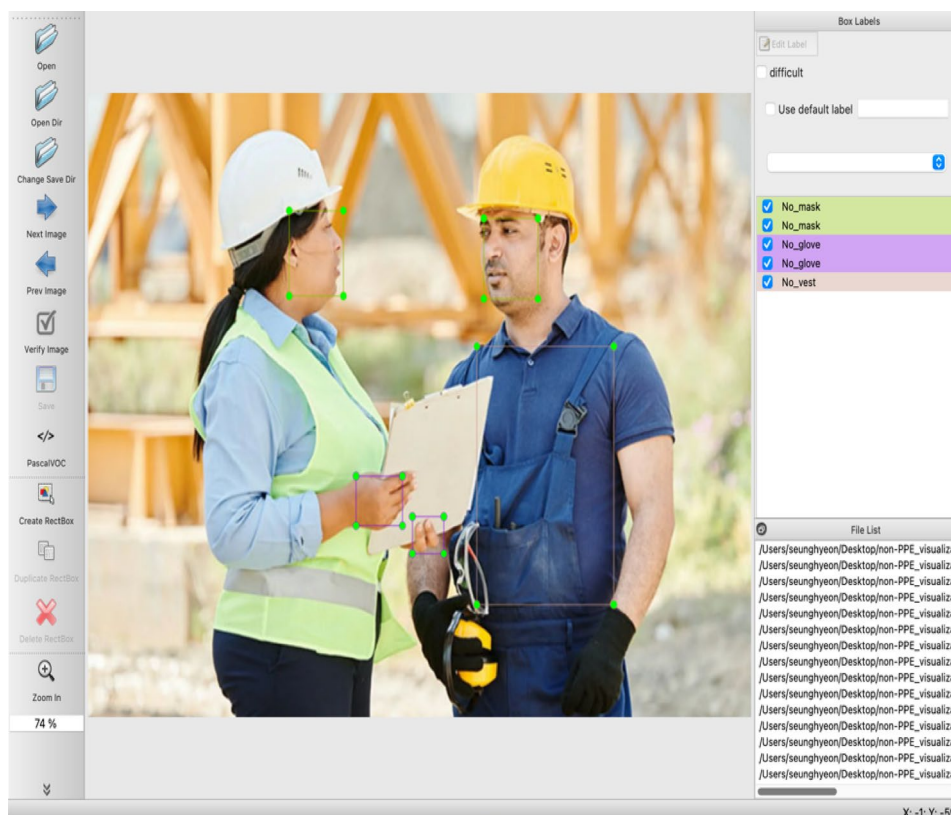


Fig. 5. Illustration of the annotation workflow for non-PPE labeling.

and evaluation. The annotation process was conducted using LabelImg, and annotations were saved in VOC XML format. An illustrative example of the annotation workflow is shown in Fig. 5.

Data split

The collected dataset (4060 images) was partitioned into training, validation, and test subsets using a 60%/20%/20% split (2436/812/812 images, respectively). This protocol enables an unbiased evaluation of the final model on unseen data. However, because some violation categories (e.g., no helmet and no vest) are relatively less frequent, a naïve random split can yield uneven class representation across subsets. To mitigate this risk, stratified sampling was employed to preserve the overall class distribution in each split, thereby reducing sampling bias and improving the reliability of both training and evaluation³⁶.

Table 2 summarizes the resulting distribution. In addition to image counts, the table reports the number of annotated violation instances per class. Note that instance counts can exceed the number of images because a single image may contain multiple workers and multiple violation types. As shown, the class proportions are consistent across training, validation, and test subsets, confirming that the stratification procedure maintained balanced class representation.

Experimental settings

All experiments were conducted on a workstation running Windows 10. The system featured a high-performance Intel Core i7-13700H CPU equipped with 14 cores and 20 threads, along with 128 GB of RAM. Computational tasks were accelerated using a powerful NVIDIA GeForce RTX 4090 GPU.

Purpose	Training	Validation	Test	Total
Images	2436 (60%)	812 (20%)	812 (20%)	4060 (100%)
No-helmet	3314 (15.63%)	1105 (15.64%)	1105 (15.63%)	5524 (15.63%)
No-mask	5701 (26.89%)	1900 (26.89%)	1901 (26.89%)	9502 (26.89%)
No-vest	3153 (14.87%)	1051 (14.87%)	1051 (14.87%)	5255 (14.87%)
No-glove	4915 (23.18%)	1638 (23.18%)	1639 (23.19%)	8192 (23.18%)
No-shoes	4117 (19.42%)	1372 (19.42%)	1373 (19.42%)	6862 (19.42%)
Total instances	21,200 (100%)	7066 (100%)	7069 (100%)	35,335 (100%)

Table 2. Class-wise dataset composition and split distribution.

Architecture	Min	25%	Median	Mean	75%	Max	Std
AP50							
Swin transformer	89.01	90.86	91.49	91.34	92.02	92.29	0.93
ViT	84.32	86.43	87.08	86.98	87.61	89.29	0.88
PVT	88.79	90.02	90.46	90.42	90.86	92.19	0.71
MobileViT	85.43	87.05	87.52	87.44	87.96	89.14	0.8
Axial transformer	88.02	89.55	90.02	89.93	90.41	91.52	0.77
Proposed method	89.56	91.92	92.44	92.36	92.93	93.78	0.86
AP50:95							
Swin transformer	61.43	65.86	66.98	66.84	67.96	68.67	1.92
ViT	55.63	58.79	59.63	59.51	60.43	62.84	1.64
PVT	63.28	66.28	67.38	67.25	68.42	69.93	1.73
MobileViT	58.88	61.09	61.92	61.86	62.61	64.27	1.27
Axial transformer	61.18	64.44	65.44	65.36	66.29	67.84	1.5
Proposed method	63.47	68.03	69.18	69.09	70.01	71.62	1.71

Table 3. Summary of detection performance (AP50 and AP50:95).

Transfer learning with a pre-trained model

Training deep detectors from random initialization typically requires large annotated datasets, substantial computational resources, and long optimization times. To mitigate these requirements, transfer learning is applied: each model is initialized with weights pretrained on a large-scale benchmark and then fine-tuned on the target dataset³⁶. COCO-pretrained weights are adopted for initialization because COCO is a widely used object-detection benchmark and provides strong supervision for both classification and bounding-box regression^{37–39}.

Pretraining on diverse object categories and scene contexts yields generalizable visual features that can accelerate convergence and improve generalization when adapting to non-PPE detection, where targets may be small, densely distributed, and embedded in cluttered backgrounds. During fine-tuning, the pretrained parameters are updated using the experimental dataset to specialize the learned representations for the target domain.

Hyperparameter optimization

Detection accuracy is highly sensitive to training hyperparameters, and performance can vary substantially under different optimization settings. Since no single configuration is optimal across tasks and datasets, hyperparameter tuning is typically conducted by specifying plausible ranges informed by prior studies and empirical evidence from related detection problems. Accordingly, the search ranges adopted here follow commonly reported YOLO-style training practices, including those used in construction-site vision applications.

For consistency across models, all methods were trained and evaluated using an identical search space and the same training protocol. Each method was tested under 243 hyperparameter combinations, and the best-performing configuration was selected for reporting. The search considered batch size {1, 2, 4}, learning rate {0.00025, 0.0005, 0.001}, weight decay {0.00001, 0.00003, 0.00005}, momentum {0.3, 0.5, 0.7}, and maximum training iterations {10,000, 20,000, 30,000}. Additional details on the chosen ranges and combinations are provided in^{40–42}.

Results and discussion

Training and validation performance

Effect of hyperparameters

Table 3 reports summary statistics for AP50 and AP50:95 across architectural variants, and Fig. 6 visualizes these distributions via box plots. A key observation is that the proposed method achieves the highest central performance on both metrics (median AP50 92.44, median AP50:95 69.18) while keeping dispersion moderate

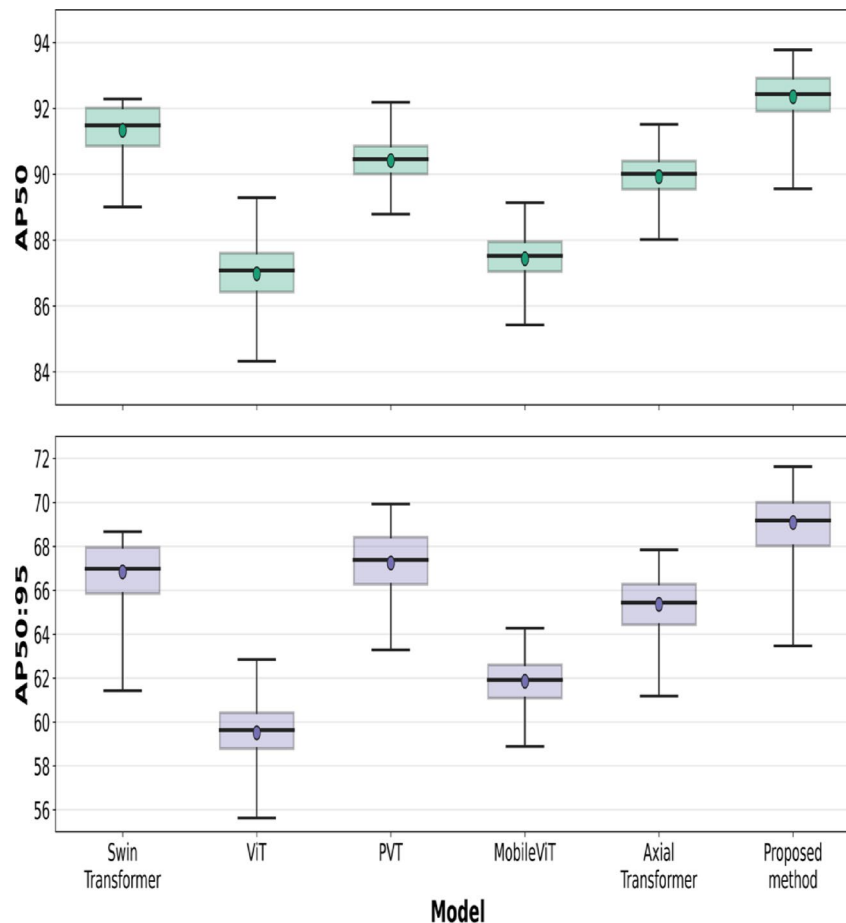


Fig. 6. Boxplots of AP50 and AP50:95 across methods.

(std 0.86 and 1.71, respectively), indicating that the gain is not driven by a few outlier runs. Another interesting pattern is the metric-dependent sensitivity: for several models, AP50 spreads are relatively tight (std ≈ 0.71 – 0.93), whereas AP50:95 spreads are noticeably larger (std ≈ 1.27 – 1.92), suggesting that stricter localization (IoU 0.50–0.95) amplifies performance variability. Among baselines, PVT shows the smallest AP50 variance (std 0.71), implying stable detection at IoU = 0.5, yet it does not translate into the top AP50:95 median. Swin Transformer remains competitive, but the proposed method still shifts both the median and upper quartile upward, reflecting consistent improvements across the performance distribution. Finally, ViT is clearly separated with the lowest medians (AP50 87.08, AP50:95 59.63), highlighting a substantial gap in both accuracy and localization quality.

Sensitivity to randomness

To assess performance variability arising from randomness in training and evaluation, each architecture was re-trained with its optimal hyperparameter configuration across ten independent weight-initialization seeds, and ten different split seeds. The resulting statistics are summarized in Table 4 as mean $\pm$ standard deviation ($\mu \pm \sigma$), and the corresponding distributions are visualized using violin plots in Fig. 8.

Sensitivity to weight initialization Table 4 and Fig. 7 indicate that, with a fixed data split, YOLOv10 performance is only weakly sensitive to random weight initialization. Across architectural variants, the standard deviations remain small (on the order of a few tenths), suggesting reliable convergence across seeds. The proposed method attains the highest mean performance while preserving low dispersion (Precision 96.02 ± 0.20 , Recall 90.18 ± 0.22 , AP50 94.73 ± 0.28 , AP50:95 70.68 ± 0.30) and exhibits a compact distribution in the violin plot. In contrast, ViT shows comparatively larger seed-dependent variation (e.g., Precision ± 0.29 , AP50:95 ± 0.34), consistent with a less stable optimization behavior.

An additional observation is that the relative ordering of architectures is stable across seeds: backbones that perform strongly on average (e.g., Proposed, PVT/ConvNeXt) remain consistently competitive, whereas weaker backbones (e.g., ViT/MobileViT) do not close the gap under different initializations. Moreover, variability is typically slightly larger for AP50:95 than AP50, implying that initialization stochasticity affects localization accuracy more than detection at the looser IoU threshold.

Sensitivity to data split When randomness is introduced through dataset partitioning, variability increases markedly (Table 4; Fig. 7), resulting in broader violin distributions than under weight initialization. Standard deviations in the split-seed setting are consistently larger across metrics, indicating that changes in the composition of training/validation/test subsets dominate performance uncertainty. Despite this, the proposed

Random factor	Architecture	Precision	Recall	AP50	AP50:95
Weight initialization	CSPNet	94.21 $\pm$ 0.16	88.31 $\pm$ 0.20	92.41 $\pm$ 0.22	68.63 $\pm$ 0.25
	ConvNeXt	94.58 $\pm$ 0.18	88.70 $\pm$ 0.22	92.93 $\pm$ 0.25	69.37 $\pm$ 0.27
	EfficientNet	94.33 $\pm$ 0.15	88.42 $\pm$ 0.19	92.56 $\pm$ 0.23	68.20 $\pm$ 0.24
	Swin Transformer	95.06 $\pm$ 0.17	89.46 $\pm$ 0.19	93.15 $\pm$ 0.26	68.60 $\pm$ 0.28
	ViT	91.41 $\pm$ 0.29	86.21 $\pm$ 0.27	89.96 $\pm$ 0.23	62.33 $\pm$ 0.34
	PVT	94.66 $\pm$ 0.17	88.78 $\pm$ 0.18	93.00 $\pm$ 0.25	69.63 $\pm$ 0.28
	MobileViT	91.62 $\pm$ 0.24	86.38 $\pm$ 0.25	89.92 $\pm$ 0.23	63.68 $\pm$ 0.30
	Axial Transformer	94.02 $\pm$ 0.19	88.12 $\pm$ 0.23	92.21 $\pm$ 0.24	67.72 $\pm$ 0.26
	Proposed method	96.02 $\pm$ 0.20	90.18 $\pm$ 0.22	94.73 $\pm$ 0.28	70.68 $\pm$ 0.30
Dataset split	CSPNet	93.84 $\pm$ 0.43	87.81 $\pm$ 0.50	92.10 $\pm$ 0.44	68.03 $\pm$ 0.63
	ConvNeXt	94.18 $\pm$ 0.44	88.20 $\pm$ 0.52	92.38 $\pm$ 0.47	68.63 $\pm$ 0.68
	EfficientNet	93.90 $\pm$ 0.40	87.96 $\pm$ 0.47	91.96 $\pm$ 0.43	67.49 $\pm$ 0.58
	Swin Transformer	95.10 $\pm$ 0.41	89.34 $\pm$ 0.45	92.62 $\pm$ 0.30	68.40 $\pm$ 0.55
	ViT	90.81 $\pm$ 0.63	85.42 $\pm$ 0.71	89.11 $\pm$ 0.64	61.54 $\pm$ 0.88
	PVT	94.10 $\pm$ 0.46	88.29 $\pm$ 0.49	92.41 $\pm$ 0.55	68.86 $\pm$ 0.61
	MobileViT	91.26 $\pm$ 0.51	85.83 $\pm$ 0.60	89.06 $\pm$ 0.54	62.94 $\pm$ 0.69
	Axial Transformer	93.62 $\pm$ 0.45	87.66 $\pm$ 0.56	91.53 $\pm$ 0.48	66.94 $\pm$ 0.67
	Proposed method	95.73 $\pm$ 0.38	90.02 $\pm$ 0.44	94.21 $\pm$ 0.41	70.02 $\pm$ 0.57

Table 4. YOLOv10 results across architectural variants over ten seeds (mean $\pm$ std).

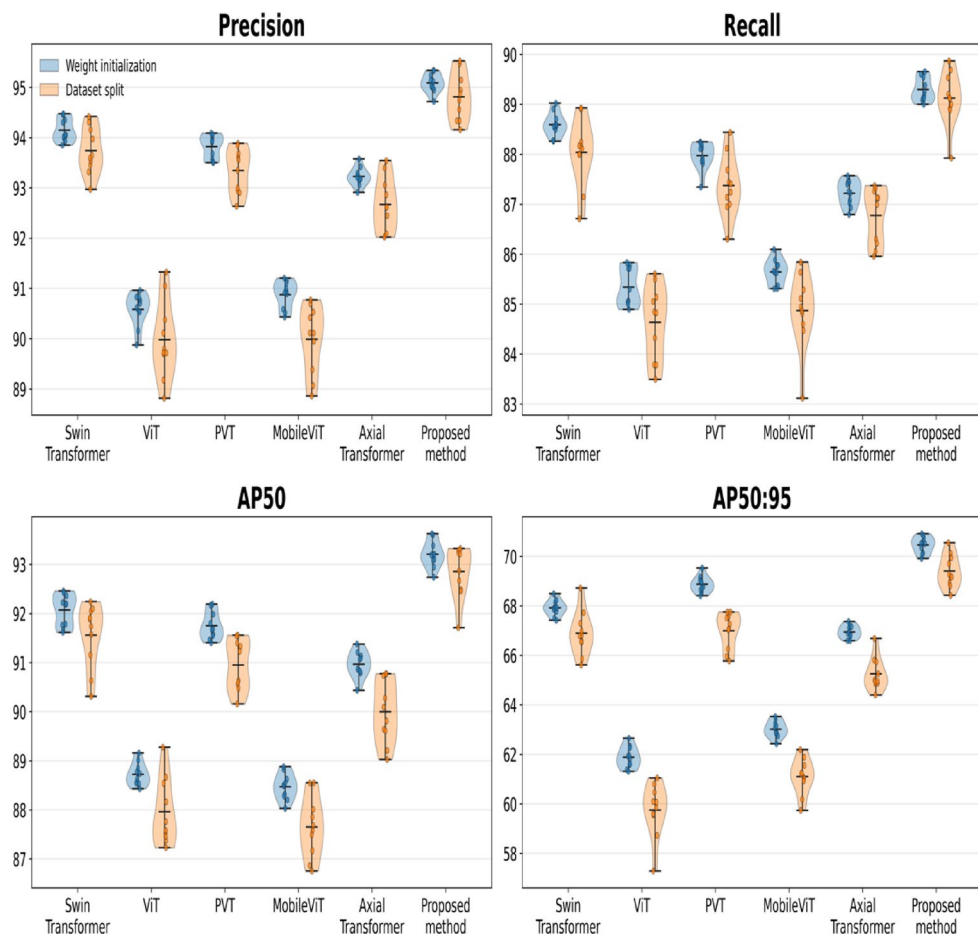


Fig. 7. Violin plots showing performance variability across 10 seeds under two randomness sources.

method remains both the best-performing and among the most stable models ($AP_{50} 94.21 \pm 0.41$, $AP_{50:95} 70.02 \pm 0.57$), evidencing robust generalization across alternative partitions. ViT again exhibits the strongest sensitivity to split selection ($AP_{50:95} 61.54 \pm 0.88$), aligning with the widest distributions.

A further noteworthy pattern is that split variation tends to penalize $AP_{50:95}$ more than AP_{50} , suggesting that partition-induced distribution shift primarily degrades bounding-box tightness (high-IoU localization) rather than coarse detection. Finally, the proposed method shows only a modest reduction in mean performance from the initialization setting to the split setting (e.g., $AP_{50:95} 70.68 \rightarrow 70.02$), indicating that its gains are not specific to a single partition and transfer across multiple splits.

Best model analysis in each method

Figure 8 summarizes the best validation run obtained for each architecture and highlights both overall performance and the accuracy–speed trade-off. The row-normalized heatmap shows that the proposed method dominates most performance criteria: it achieves the highest mAP (AP_{50}) (93.78) and mAP ($AP_{50:95}$) (71.62), and it also leads consistently at the class level (e.g., AP_{50} for No-helmet 95.12, No-mask 94.36, and corresponding $AP_{50:95}$ values 73.05 and 72.14). Importantly, the proposed method maintains advantages across object scales, yielding the best results for AP_{ES} (85.92), AP_S (91.18), AP_M (94.72), and AP_L (96.63), indicating improved detection not only for large objects but also for extremely small targets.

The scatter plot reinforces this conclusion: the proposed method sits at the top of the accuracy axis, confirming the best detection quality, while retaining competitive inference speed (57.28 FPS). In contrast, MobileViT is the fastest (80.65 FPS) but shows a clear accuracy gap (mAP(AP_{50}) 89.14), illustrating a throughput–accuracy trade-off. Strong baselines such as PVT and Swin Transformer are relatively accurate but remain below the proposed method on both AP_{50} and $AP_{50:95}$.

Test set evaluation

Validation-to-test generalization

Figure 9 summarizes the generalization performance of the proposed method from the validation set to the test set. Overall, the model exhibits consistent but moderate performance degradation across metrics, suggesting stable generalization rather than overfitting to the validation data. On the test set, the proposed method achieves 92.49 mAP@50 and 70.04 mAP@50:95, corresponding to validation-to-test gaps of – 1.52 and – 1.72, respectively.

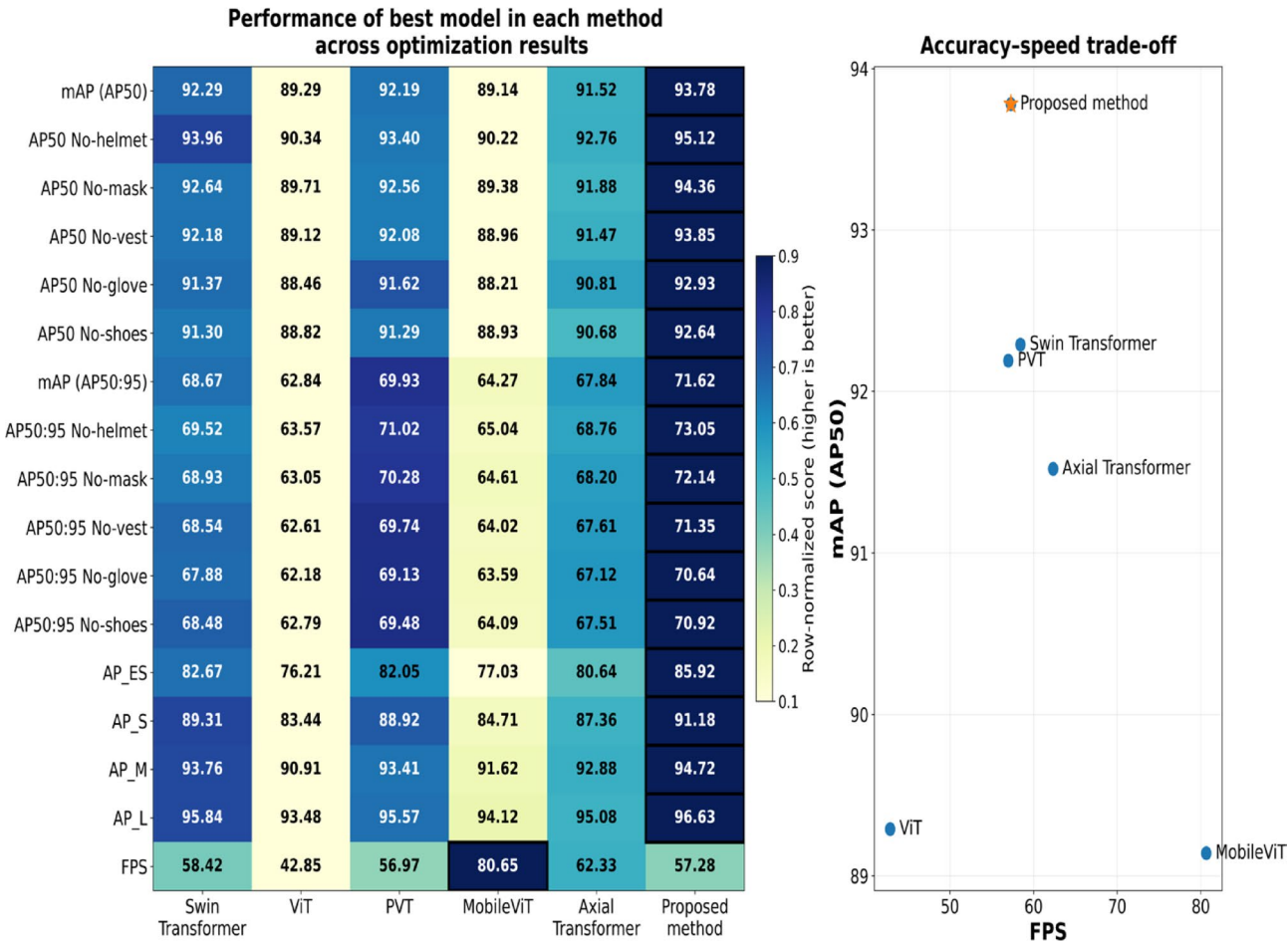


Fig. 8. Best-run selection per method based on the highest validation score.

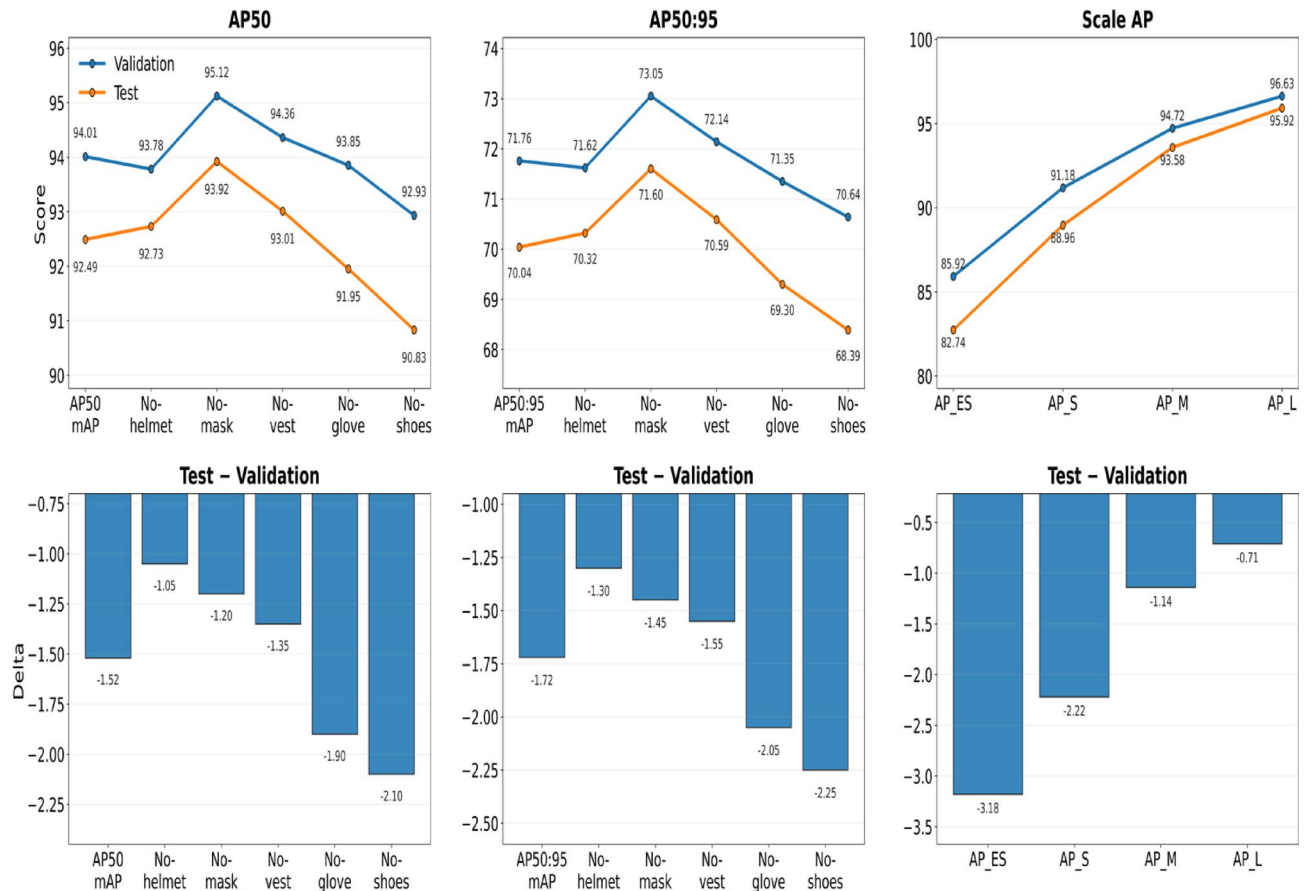


Fig. 9. Validation-to-test generalization performance comparison.

At the class level, AP50 decreases are limited to -1.05 to -2.10 , with the largest drops observed for No-shoes (-2.10) and No-glove (-1.90), indicating that these categories remain the most challenging under distribution shift. A similar trend is observed for AP50:95, where the gaps range from -1.30 to -2.25 , again peaking for No-shoes (-2.25) and No-glove (-2.05). This suggests that stricter localization requirements amplify the difficulty for these classes.

Scale-based results further show that the largest generalization losses occur for extremely small objects (AP_{ES} : -3.18) and small objects (AP_S : -2.22), whereas performance is comparatively stable for medium (AP_M : -1.14) and large (AP_L : -0.71) objects. Overall, the primary generalization bottlenecks arise from tiny/small targets and fine-grained categories, while detection of larger targets transfers reliably to the test set.

Effect of image conditions

Test-set performance was further analyzed by stratifying the test data according to six image-condition factors (Table 5), each of which splits the test set into two subsets. Representative detection examples are provided in Fig. 10. The performance consistently degrades under more challenging conditions, and the reduction is generally larger for mAP@50:95 than for mAP@50, indicating that adverse conditions primarily impair localization tightness rather than coarse detection. Among the factors, visual range and worker overlap have the strongest impact. Moving from near-field to far-field views reduces mAP@50 from 93.42 to 91.22 (-2.20) and mAP@50:95 from 71.03 to 68.71 (-2.32), with a pronounced loss for extremely small objects (AP_{ES} : 85.12 to 78.46, -6.66). Similarly, overlapping workers cause a substantial drop from 93.12 to 90.94 (-2.18) at mAP@50 and from 70.77 to 68.27 (-2.50) at mAP@50:95, accompanied by the largest degradation for tiny targets (AP_{ES} : 84.22 to 77.12, -7.10).

Illumination also matters: low illumination decreases mAP@50:95 by -2.14 and AP_{ES} by -4.18 . In contrast, the indoor/outdoor split does not show degradation for outdoor scenes in this dataset, with slightly higher mAP values outdoors. Overall, these results indicate that the proposed method remains robust across diverse site conditions, while its main limitations arise in far-field and occluded/crowded scenarios, where tiny instances are hardest to localize accurately.

Ablation study

Component-wise ablation of the proposed method Table 6 presents the component-wise ablation of ODConv, EMA-BiFPN, MPDIoU, and the four-scale prediction head on YOLOv10. Variant 1 is the baseline ($AP50 = 90.71$, $AP50:95 = 67.40$, $FPS = 63.47$). Variants 2–5 test each component individually, Variants 6–8 evaluate selected

Image condition	Illumination		Visual range		Space		Overlap of worker		Density of worker		Number of workers	
Range	High	Low	Near-field view	Far-field view	Indoor	Outdoor	Non-overlap	Overlap	Scatter	Grouped	Single	Multiple
mAP (AP50)	93.21	91.3	93.42	91.22	92.08	92.67	93.12	90.94	92.96	91.46	93.28	91.58
No-helmet	93.28	91.62	93.85	91.24	92.61	92.91	93.34	91.35	93.05	91.88	93.57	91.68
No-mask	94.55	92.7	94.62	92.85	93.66	94.14	94.62	92.8	94.28	93.12	94.88	92.94
No-vest	93.62	92.01	93.78	92.26	92.94	93.18	93.72	92.16	93.39	92.53	93.96	92.37
No-glove	92.76	90.68	92.88	90.77	91.6	92.11	92.64	89.96	92.26	90.79	92.86	90.36
No-shoes	91.84	89.49	91.96	88.96	89.59	91.01	91.28	88.44	91.82	89	91.13	90.55
mAP (AP50:95)	70.83	68.69	71.03	68.71	69.53	70.32	70.77	68.27	70.44	68.95	70.91	69.26
No-helmet	71.05	69.12	71.18	68.86	69.86	70.54	70.98	68.74	70.72	69.1	71.02	69.42
No-mask	72.36	70.05	72.55	70.42	71.05	71.88	72.31	70.38	72.05	70.83	72.46	70.96
No-vest	71.34	69.38	71.62	69.58	70.14	70.92	71.42	69.44	71.1	69.98	71.55	70.08
No-glove	70.2	67.95	70.36	67.98	68.72	69.56	70.1	66.92	69.72	67.88	70.21	68.21
No-shoes	69.18	66.94	69.44	66.73	67.88	68.68	69.05	65.88	68.62	66.98	69.32	67.63
AP_{ES}	84.06	79.88	85.12	78.46	81.26	83.38	84.22	77.12	83.64	79.26	84.86	79.92
AP_S	89.78	87.62	90.02	86.44	88.12	89.24	90.01	86.38	89.72	87.48	90.34	87.66
AP_M	94.05	92.84	94.02	92.95	93.24	93.78	94.08	92.61	93.92	93.02	94.24	93.12
AP_L	96.18	95.31	96.24	95.4	95.66	96.06	96.28	95.01	96.14	95.39	96.32	95.53

Table 5. Test-set performance stratified by image condition.

two-module combinations, Variant 9 represents the three-module configuration without the four-scale head, and Variant 10 is the full model.

Among the single-module variants, ODConv (Variant 2) and EMA-BiFPN (Variant 3) produce similar AP50 gains (+0.73 and +0.74), with EMA-BiFPN showing slightly stronger improvements for AP_{ES} and AP_S , indicating a greater benefit for small-object feature fusion. In contrast, MPDIoU (Variant 4) yields the largest AP50:95 gain (+1.06) while preserving baseline speed, suggesting that its main role is improved localization rather than stronger coarse detection. The four-scale prediction head (Variant 5) gives a smaller overall AP increase, but its effect is more concentrated on extremely small and small targets (AP_{ES} +3.04; AP_S +1.29), highlighting its specific value for tiny PPE-related cues.

Among the two-module settings, ODConv + EMA-BiFPN (Variant 6) achieves the highest AP50 (92.02) and the strongest AP_{ES} gain (+3.68), but at the largest speed cost. ODConv + MPDIoU (Variant 7) and EMA-BiFPN + MPDIoU (Variant 8) provide more balanced trade-offs, with Variant 8 achieving the best AP50:95 among the two-module combinations.

A particularly important observation comes from Variants 9 and 10. Adding the four-scale head to the three-module configuration improves AP50 from 92.18 to 92.49 and AP50:95 from 69.08 to 70.04, while AP_{ES} rises from 81.12 to 82.74. This shows that the four-scale design remains complementary rather than redundant. Overall, ODConv and EMA-BiFPN mainly enhance representation and multi-scale fusion, MPDIoU improves localization quality, and the four-scale head most clearly strengthens tiny-object detection.

Comparison with other detectors

Comparative detectors and implementation settings To evaluate the proposed approach under consistent experimental conditions, comparative experiments were conducted using representative recent one-stage, two-stage, and transformer-based object detectors. First, YOLOv10 was used as a common detection framework while varying only the backbone network, including CNN-based backbones (CSPNet, ConvNeXt, and EfficientNet) and transformer-based backbones (Swin Transformer, Vision Transformer, Pyramid Vision Transformer, MobileViT, and Axial Transformer). With the detection head and training pipeline held constant, the resulting performance differences primarily reflect the influence of backbone design^{43–45}.

To broaden the benchmarking scope, additional comparisons were conducted with recent external detectors. The one-stage group included YOLOv8 variants (CSPDarknet-DETR, CSPDarknet-T-FPN, and CSPDarknet-CBAM) and YOLOv9 models at multiple scales (YOLOv9t–YOLOv9e). The transformer-based group included DINO with ResNet-50 and Swin Transformer backbones, as well as Deformable DETR with ResNet-50 and ResNet-101 backbones. The two-stage group comprised Faster R-CNN with FPN using ResNet-101, ResNet-152, and MobileNetV3 backbones. To ensure fair comparison, the baseline detectors were implemented and trained using the MMDetection framework⁴⁶, whereas the YOLOv8-based baselines were built and trained using the Ultralytics toolkit⁴⁷. Table 7 summarizes the comparative detectors and architectural settings.

Comparison results Table 8 compares the proposed method against two-stage detectors (Faster R-CNN), classical one-stage detectors (SSD500 and R-FCN), recent YOLO variants (YOLOv8/YOLOv9), and transformer-based detectors (Deformable DETR and DINO) for PPE (helmet-related) monitoring. Overall, the proposed method achieves the best test-set accuracy, reaching AP50 = 92.49 and AP50:95 = 70.04, while maintaining real-time throughput (57.28 FPS).

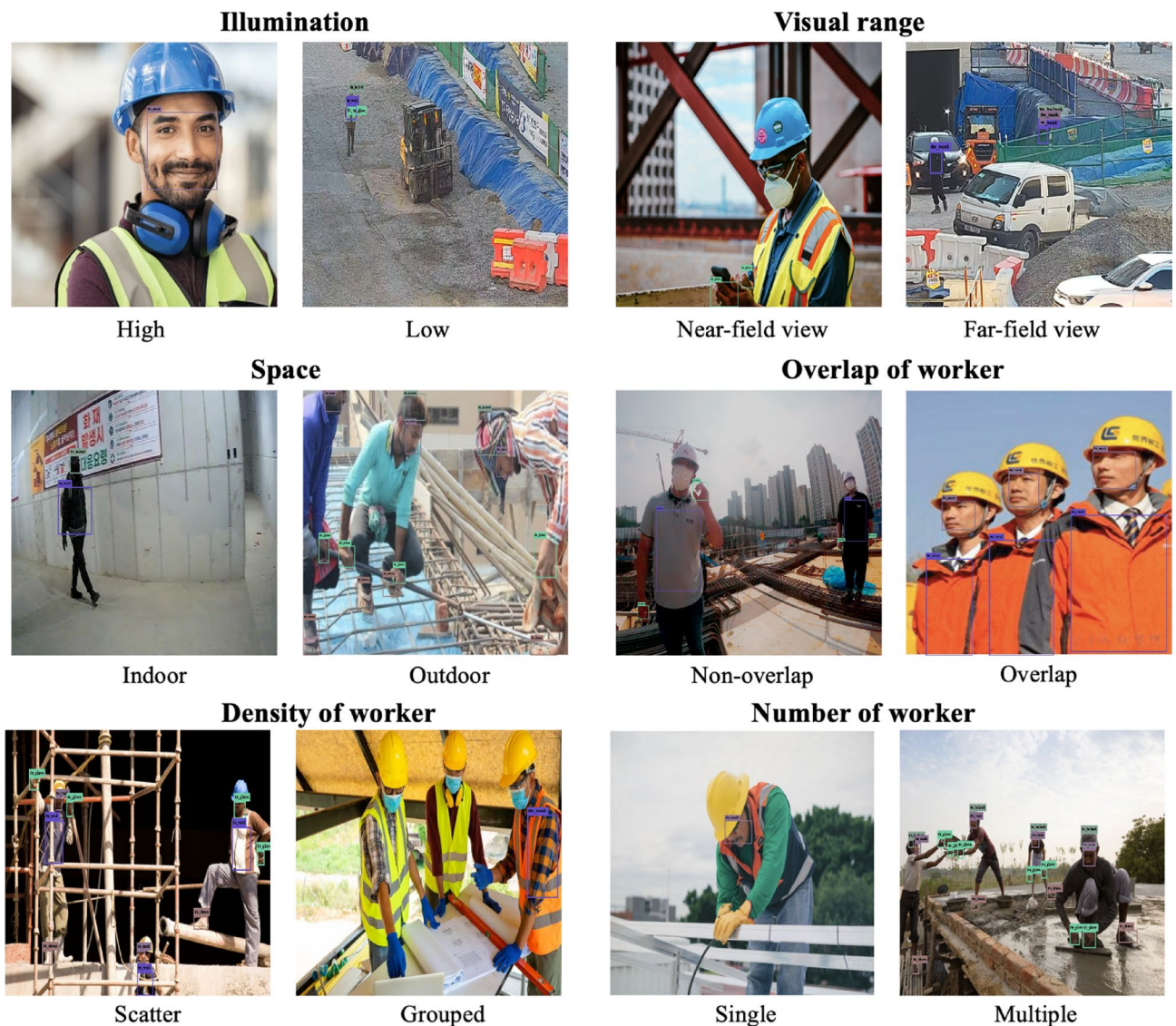


Fig. 10. Qualitative detection examples under diverse scene conditions.

Among the baselines, YOLOv9e is the strongest accuracy competitor (AP_{50} 91.35, $AP_{50:95}$ 68.72), but it remains inferior to the proposed method by +1.14 AP_{50} and +1.32 $AP_{50:95}$, and it is also slower (42.19 FPS). Transformer-based detectors are competitive in $AP_{50:95}$ (e.g., DINO+Swin Transformer: 68.45), yet their throughput is substantially lower (18.74 FPS), limiting deployment in real-time settings. Deformable DETR variants are similarly slower (≈ 12 – 14 FPS) and do not surpass the proposed method in accuracy ($AP_{50:95} \leq 66.88$). Two-stage detectors show the largest gap, with AP_{50} around 84–86 and very low throughput (≈ 10 – 16 FPS).

The proposed method's advantage is most pronounced for small-scale targets, achieving the highest $AP_{ES} = 82.74$ and $AP_S = 88.96$. Compared with YOLOv9e, this corresponds to gains of +3.12 (AP_{ES}) and +1.00 (AP_S), indicating improved robustness for distant workers and tiny PPE instances. Improvements are smaller but consistent for medium/large objects ($AP_M = 93.58$, $AP_L = 95.92$), suggesting that the main performance differentiation arises in difficult, small-object regimes rather than easy close-range cases.

Figure 11 visualizes the accuracy–speed trade-off (AP_{50} vs FPS and $AP_{50:95}$ vs FPS). The proposed method occupies a favorable region, outperforming faster but less accurate YOLO configurations (e.g., YOLOv9t at 79.36 FPS with $AP_{50:95}$ 65.42) while providing substantially higher throughput than transformer detectors, supporting its suitability for deployment-oriented, real-time PPE monitoring on construction sites.

External validation on unseen construction sites

To further evaluate generalization to unseen construction sites, both the baseline YOLOv10 and the proposed method were additionally tested on an external dataset collected from two previously unseen construction sites. This external dataset was excluded entirely from training, validation, hyperparameter tuning, and model selection.

Model	1	2	3	4	5	6	7	8	9	10 (Full)
ODConv	–	✓	–	–	–	✓	✓	–	✓	✓
EMA-BiFPN	–	–	✓	–	–	✓	–	✓	✓	✓
MPDIoU	–	–	–	✓	–	–	✓	✓	✓	✓
Four-scale head	–	–	–	–	✓	–	–	–	–	✓
AP50 (mAP)	90.71	91.44	91.45	91.05	91.32	92.02	91.71	91.71	92.18	92.49
No-helmet	91.19	91.77	91.81	91.48	91.56	92.27	92	92.03	92.34	92.73
No-mask	92.21	92.87	92.92	92.52	92.71	93.44	93.13	93.15	93.53	93.92
No-vest	91.49	92.1	92.18	91.82	91.96	92.66	92.37	92.44	92.72	93.01
No-glove	90.02	90.86	90.79	90.38	90.62	91.47	91.15	91.07	91.56	91.95
No-shoes	88.65	89.58	89.53	89.06	89.33	90.28	89.91	89.85	90.36	90.83
AP50:95 (mAP)	67.4	67.77	67.8	68.46	67.88	68.05	68.65	68.67	69.08	70.04
No-helmet	67.85	68.17	68.2	68.87	68.31	68.42	69.03	69.05	69.42	70.32
No-mask	68.97	69.38	69.43	70.05	69.51	69.72	70.28	70.31	70.74	71.6
No-vest	68.08	68.44	68.48	69.13	68.62	68.73	69.32	69.35	69.78	70.59
No-glove	66.59	67.04	67.01	67.75	67.16	67.34	68	67.97	68.41	69.3
No-shoes	65.53	65.8	65.86	66.51	65.81	66.04	66.63	66.68	67.12	68.39

Table 6. Ablation results of the proposed components.

Detector family	Detector	Architecture
One-stage detector	YOLOv10	CSPNet / ConvNeXt / EfficientNet / Swin Transformer / ViT / PVT / MobileViT / Axial Transformer
	YOLOv9	YOLOv9t / YOLOv9s / YOLOv9m / YOLOv9c / YOLOv9e
	YOLOv8	CSPDarknet-DETR / CSPDarknet-T-FPN / CSPDarknet-CBAM
Two-stage detector	Faster R-CNN + FPN	ResNet-101 / ResNet-152 / MobileNetV3
Transformer detector	Deformable DETR	ResNet-50 / ResNet-101
	DINO	ResNet-50 / Swin Transformer

Table 7. Comparison of detection frameworks and key architectural modules.

The external test set consisted of 313 images and 2937 annotated non-compliance instances, including 573 no-helmet, 687 no-mask, 476 no-vest, 612 no-glove, and 589 no-shoes cases. Because obtaining sufficient no-shoes and no-vest examples directly from real construction-site imagery was difficult, additional images from external sources were incorporated to ensure adequate representation across all five non-compliance categories. The composition of the external test set is summarized in Table 9.

Comparison on the internal and external test sets Table 10 compares the performance of the baseline YOLOv10 and the proposed method on the internal and external test sets and additionally reports the performance variation. As shown in Table 10, both models exhibited performance reductions on the external test set, confirming that the unseen-site setting was more challenging than the internal test set. For the baseline YOLOv10, AP50 and AP50:95 decreased by 0.77 and 1.12, respectively, while AP_{ES} and AP_S decreased by 1.18 and 0.84. The proposed method also showed performance drops on the external test set, but the reductions were slightly smaller across all key metrics, with decreases of 0.61 in AP50, 0.88 in AP50:95, 1.06 in AP_{ES} , and 0.77 in AP_S .

The proposed method also consistently outperformed the baseline on both datasets. On the external test set, it achieved AP50=91.88 and AP50:95=69.16, compared with 89.94 and 66.28 for the baseline, corresponding to gains of 1.94 and 2.88, respectively. The improvement was particularly evident for extremely small and small PPE-related targets, where the proposed method exceeded the baseline by 6.02 in AP_{ES} and 2.11 in AP_S . In contrast, the gains for medium- and large-scale targets were smaller (+1.27 in AP_M and +0.76 in AP_L).

At the class level, the external test set generally caused larger degradation for no-glove and no-shoes than for no-helmet and no-mask, especially under AP50:95. This trend is consistent with the weaker and smaller visual cues typically associated with hand- and foot-related PPE violations. Nevertheless, the proposed method remained superior to the baseline across all five non-compliance categories on both AP50 and AP50:95.

Confusion matrix analysis on the internal and external test sets To further examine class-level error patterns, confusion matrices were generated for the proposed method on the internal and external test sets, as shown in Fig. 12. The top row presents the confusion matrices in raw counts, whereas the bottom row shows the corresponding row-normalized percentages. In both settings, the diagonal entries remain dominant, indicating that most PPE violation instances were correctly classified.

The count-based matrices show that the overall classification structure is preserved across both datasets, but the external test set exhibits slightly greater off-diagonal confusion than the internal test set. This pattern indicates that unseen-site conditions increase the difficulty of class discrimination. The most noticeable confusion occurs

Detector	Architecture	AP50 (mAP)	No-helmet	No-mask	No-vest	No-glove	No-shoes	AP50:95 (mAP)	No-helmet	No-mask	No-vest	No-glove	No-shoes	AP _{ES}	AP _S	AP _M	AP _L	FPS
Faster R-CNN	ResNet-101	85.92	86.41	86.94	86.32	85.52	84.41	60.74	61.1	61.64	60.83	60.29	59.83	71.48	80.92	88.63	92.47	10.62
	ResNet-152	86.45	86.83	87.58	86.65	86.05	85.14	61.23	61.55	62.3	61.34	60.68	60.27	72.05	81.47	89.12	92.88	9.84
	MobileNetV3	84.18	84.78	85.49	84.56	83.89	82.18	58.92	59.44	59.71	59.3	58.39	57.76	68.96	78.84	87.02	91.45	16.35
SSD500	ResNet-101	83.42	83.84	84.43	83.7	83.28	81.84	57.63	57.87	58.61	57.93	57.13	56.61	66.57	76.98	85.89	90.94	26.18
	ResNet-152	83.97	84.54	84.96	84.16	83.61	82.57	58.01	58.42	58.93	58.19	57.59	56.91	67.05	77.44	86.31	91.26	24.77
	MobileNetV3	81.64	82.17	82.9	81.77	81.18	80.19	55.92	56.43	56.91	56.14	55.55	54.58	63.74	74.52	84.22	89.85	33.46
R-FCN	ResNet-101	84.96	85.54	86.05	85.34	84.62	83.25	59.72	60.09	60.72	59.87	59.4	58.52	69.35	79.22	87.61	91.83	13.06
	ResNet-152	85.38	85.88	86.63	85.67	84.93	83.79	60.11	60.43	61.06	60.2	59.79	59.07	69.92	79.66	88.05	92.18	12.48
	MobileNetV3	83.25	83.69	84.5	83.54	82.85	81.67	57.86	58.2	58.79	58.07	57.41	56.83	66.08	76.35	86.02	90.62	19.92
YOLOv8	CSPDarknet-DETR	89.56	89.92	90.86	89.91	89.16	87.95	66.42	66.78	67.58	66.79	65.97	64.98	77.12	85.88	91.96	94.72	52.73
	CSPDarknet-T-FPN	88.76	89.4	89.94	89.21	88.32	86.93	65.74	66.2	66.65	65.99	65.2	64.66	76.34	85.22	91.48	94.39	73.58
	CSPDarknet-CBAM	89.18	89.62	90.46	89.53	88.82	87.47	66.1	66.54	67.16	66.41	65.55	64.84	76.92	85.61	91.72	94.55	71.24
YOLOv9	YOLOv9t	88.54	89.04	89.63	88.83	88.04	86.16	65.42	65.77	66.43	65.56	64.86	64.46	76.18	84.96	91.22	94.06	79.36
	YOLOv9s	89.32	89.95	90.51	89.74	88.91	87.49	66.28	66.57	67.24	66.4	65.7	65.49	77.03	85.54	91.64	94.44	68.91
	YOLOv9m	90.21	90.87	91.45	90.63	89.78	88.32	67.12	67.47	68.12	67.26	66.48	66.27	78.25	86.82	92.35	95.06	55.47
	YOLOv9c	90.92	91.56	92.26	91.22	90.32	89.24	67.98	68.35	69.05	68.2	67.46	66.84	79.04	87.44	92.88	95.41	48.63
	YOLOv9e	91.35	91.79	92.7	91.71	90.91	89.64	68.72	69.06	69.66	68.97	68.06	67.85	79.62	87.96	93.12	95.58	42.19
Deformable DETR	ResNet-50	88.96	89.67	90.13	89.35	88.41	87.24	65.97	66.42	67.05	66.27	65.39	64.72	75.68	84.92	91.58	94.66	14.02
	ResNet-101	89.74	90.41	90.89	90	89.14	88.26	66.88	67.19	67.96	67.12	66.31	65.82	76.35	85.54	92.03	94.98	12.15
DINO	ResNet-50	90.38	90.76	91.69	90.74	90	88.71	67.92	68.29	68.7	68.18	67.64	66.79	77.44	86.63	92.52	95.18	21.36
	Swin Transformer	91.12	91.8	92.34	91.38	90.82	89.25	68.45	68.68	69.5	68.71	68.1	67.26	78.12	87.22	92.84	95.36	18.74
	92.49	Proposed method	92.73	93.92	93.01	91.95	90.83	70.04	70.32	71.6	70.59	69.3	68.39	82.74	88.96	93.58	95.92	57.28

Table 8. Comparative test performance using detector–backbone variants.

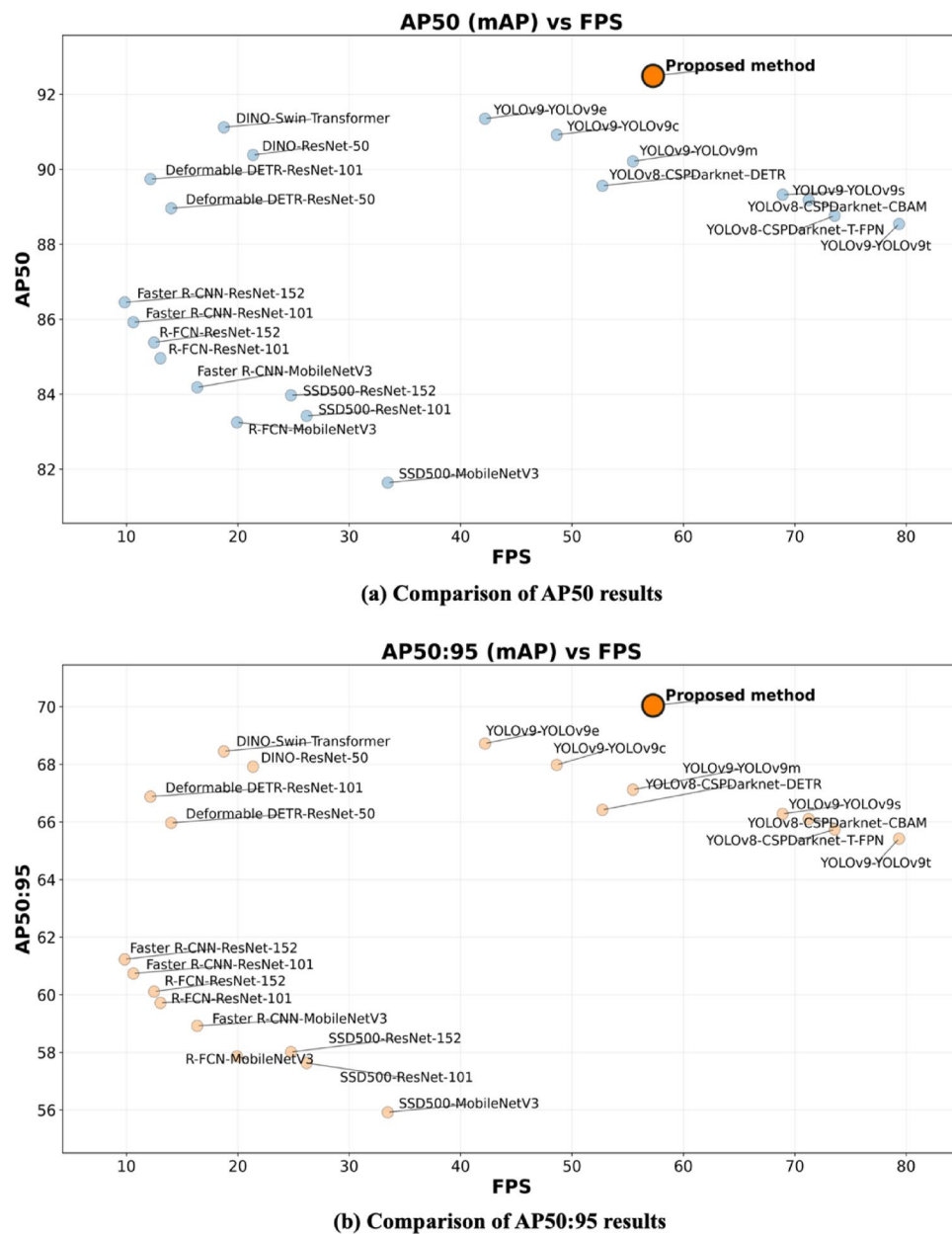


Fig. 11. Pareto analysis of accuracy and speed: (a) AP50 vs. FPS and (b) AP50:95 vs. FPS.

Purpose	Real construction sites	General images	External test
Images	132	181	313
No-helmet	235	338	573
No-mask	186	501	687
No-vest	105	371	476
No-glove	210	402	612
No-shoes	108	481	589
Total instances	844	2093	2937

Table 9. External test-set composition.

Performance	Baseline YOLOv10			Proposed method		
	Internal test	External test	Variation	Internal test	External test	Variation
AP50 (mAP)	90.71	89.94	− 0.77	92.49	91.88	− 0.61
No-helmet	91.19	90.46	− 0.73	92.73	92.12	− 0.61
No-mask	92.21	91.61	− 0.6	93.92	93.39	− 0.53
No-vest	91.49	90.78	− 0.71	93.01	92.41	− 0.6
No-glove	90.02	88.98	− 1.04	91.95	91.06	− 0.89
No-shoes	88.65	87.52	− 1.13	90.83	89.81	− 1.02
AP50:95 (mAP)	67.4	66.28	− 1.12	70.04	69.16	− 0.88
No-helmet	67.85	66.87	− 0.98	70.32	69.49	− 0.83
No-mask	68.97	68.12	− 0.85	71.6	70.88	− 0.72
No-vest	68.08	67.04	− 1.04	70.59	69.76	− 0.83
No-glove	66.59	65.18	− 1.41	69.3	68.07	− 1.23
No-shoes	65.53	63.98	− 1.55	68.39	67.05	− 1.34
AP _{ES}	76.84	75.66	− 1.18	82.74	81.68	− 1.06
AP _S	86.92	86.08	− 0.84	88.96	88.19	− 0.77
AP _M	92.41	91.87	− 0.54	93.58	93.14	− 0.44
AP _L	95.21	94.91	− 0.3	95.92	95.67	− 0.25
FPS	63.47	63.47	–	57.28	57.28	–

Table 10. Comparison results on the internal and external test sets.

between no-glove and no-shoes, which is consistent with the weaker, smaller, and more localized visual cues associated with hand- and foot-related PPE violations.

The row-normalized matrices provide a clearer view of class purity across categories. On the internal test set, the diagonal percentages are consistently high, ranging from approximately 97.9 to 98.5%. On the external test set, the diagonal remains dominant but decreases modestly to approximately 94.9 to 95.4%, confirming that class-level prediction quality is largely preserved under unseen-site conditions, although with some degradation. Among the categories, no-glove and no-shoes again show relatively stronger cross-class confusion, whereas no-helmet and no-mask remain comparatively more separable across both settings.

Visualization of extremely small targets on unseen sites To further examine the practical behavior of the proposed method in the ES regime, representative examples from the external test set are shown in Fig. 13. Each example includes the original image, the corresponding prediction result of the proposed method, and an enlarged view of the target region. As defined in Section “Size-specific average precision”, the ES category corresponds to targets with $\min(w, h) < 8$ pixels after resizing to 640×640 , meaning that the relevant visual cues often occupy only a few pixels and may therefore appear highly blurred.

The examples include both successful and unsuccessful ES detections. In some cases, the proposed method correctly identifies the ES target despite its weak visual appearance, indicating that useful PPE-related cues can still be recovered under very limited resolution. In other cases, the model fails to detect the ES target, particularly for no-mask instances, confirming that this regime remains inherently challenging. These qualitative results show that the ES regime is not uniformly reliable and that detection difficulty remains high when the visual evidence is highly blurred, weakly distinguishable, or affected by scene complexity.

Conclusions

This study presented a task-specific enhanced YOLOv10 framework for multi-category PPE non-compliance detection on construction sites by integrating ODConv, EMA-BiFPN, MPDIoU, and a four-scale prediction strategy to better handle small, weak, and partially occluded violation cues.

Experimental results on the benchmark dataset (4060 images with stratified splits and a controlled hyperparameter search) demonstrate that the proposed method achieves the best overall test performance, reaching 92.49 mAP@50 and 70.04 mAP@50:95 at 57.28 FPS. The model also yields consistently strong class-wise results, with test AP50 ranging from 90.83–93.92 across the five categories, and AP50:95 ranging from 68.39–71.60, indicating balanced performance rather than gains concentrated in a single class.

The most pronounced gains are observed under challenging scale regimes. The proposed method achieves $AP_{ES} = 82.74$ and $AP_E = 88.96$, outperforming strong baselines such as YOLOv9e by + 3.12 (AP_{ES}) and + 1.00 (AP_S), highlighting improved robustness to distant workers and tiny PPE cues. The generalization analysis indicates a moderate validation-to-test gap (e.g., − 1.52 for mAP@50 and − 1.72 for mAP@50:95), with the largest degradations observed for No-shoes and No-glove, and for extremely small objects (AP_{ES} gap − 3.18), which aligns with the most difficult visual conditions. Condition-stratified evaluation further confirms that the largest performance drops occur in far-field views and overlapping-worker scenarios, where AP_{ES} decreases most sharply, indicating that occlusion and tiny-object ambiguity remain key failure modes.

Additional external validation on two previously unseen construction sites provided a stricter assessment of robustness beyond the standard internal split. Although both the baseline YOLOv10 and the proposed method showed some performance degradation under unseen-site conditions, the proposed method remained

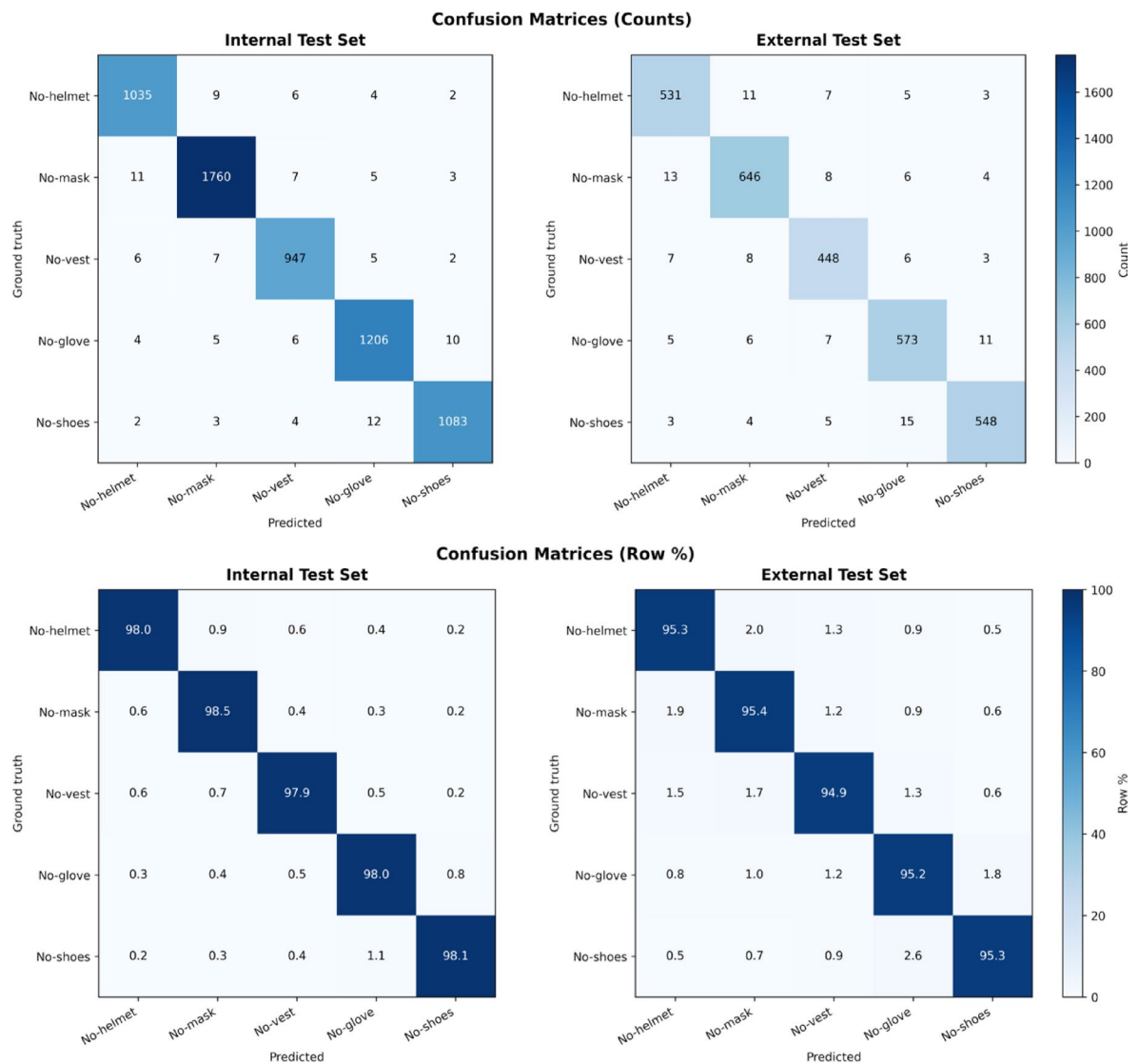
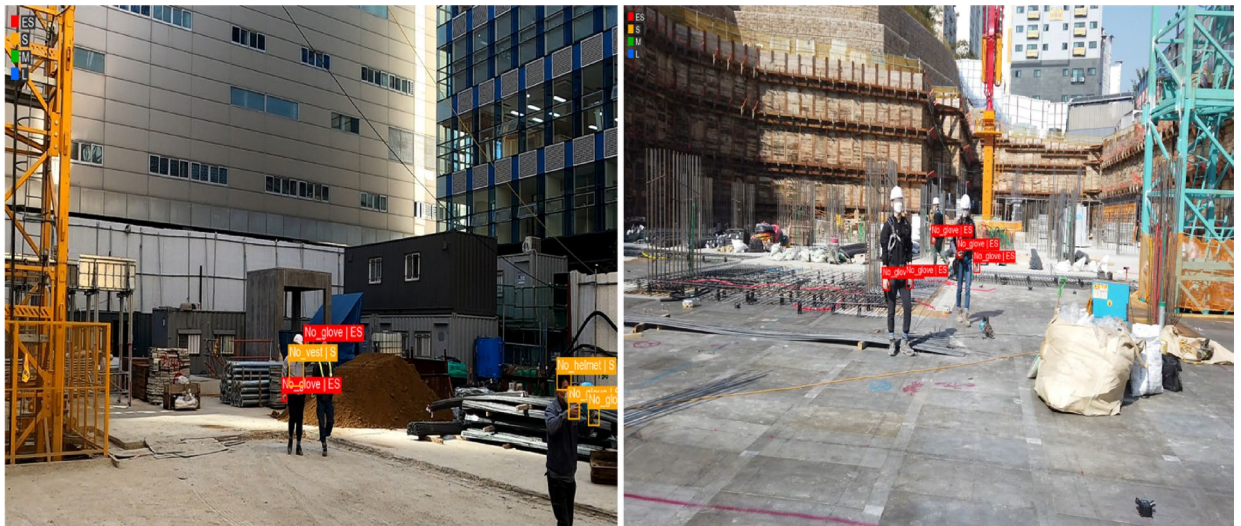


Fig. 12. Confusion matrices of the proposed method on the internal and external test sets.

consistently superior and exhibited slightly smaller reductions across key metrics, particularly for extremely small and small targets.

Overall, the results demonstrate that the proposed method offers a strong accuracy-speed trade-off for deployment-oriented PPE monitoring. Future work should prioritize improving generalization to unseen sites and camera setups, for example through site-wise splitting protocols, domain adaptation, and training strategies that explicitly target far-field and crowded scenes. In addition, deployment-focused optimization (e.g., model compression and edge acceleration) and the use of richer temporal cues from video could further enhance robustness under occlusion and motion blur, particularly for small PPE instances.



(a) Examples of correctly detected images



(b) An example of incorrectly detected images

Fig. 13. Representative examples of extremely small-target detection on the external test set using the proposed method.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 28 January 2026; Accepted: 11 May 2026

Published online: 29 May 2026

References

- Chen, S. & Demachi, K. Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. *Autom. Constr.* <https://doi.org/10.1016/j.autcon.2021.103619> (2021).
- Eum, I., Kim, J., Wang, S. & Kim, J. Heavy equipment detection on construction sites using You Only Look Once (YOLO-Version 10) with transformer architectures. *Appl. Sci.* <https://doi.org/10.3390/app15052320> (2025).
- Cha, Y. J., Choi, W., Suh, G., Mahmoudkhani, S. & Büyükoztürk, O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Comput. Aided Civ. Infrastruct. Eng.* <https://doi.org/10.1111/mice.12334> (2018).
- Hu, H., Wang, L., Zhang, M., Ding, Y. & Zhu, Q. Fast and regularized reconstruction of building facades from street-view images using binary integer programming, in: ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. (2020). <https://doi.org/10.5194/isprs-annals-V-2-2020-365-2020>.
- H. Mokayed, G. Alsayed, F. Lodin, O. Hagner, B. Backe, Enhancing Object Detection in Snowy Conditions: Evaluating YOLO v9 Models with Augmentation Techniques. In 2024 11th Int. Conf. Internet Things Syst. Manag. Secur. 198–203. (IEEE, 2024)

6. Wang, S., Eum, I., Park, S. & Kim, J. A labelled dataset for rebar counting inspection on construction sites using unmanned aerial vehicles. *Data Brief* <https://doi.org/10.1016/j.dib.2024.110720> (2024).
7. Wang, S. Development of approach to an automated acquisition of static street view images using transformer architecture for analysis of building characteristics. *Sci. Rep.* **15**, 29062. <https://doi.org/10.1038/s41598-025-14786-3> (2025).
8. Xu, Z., Jha, N., Mehadi, S. & Mandal, M. Multiscale object detection on complex architectural floor plans. *Autom. Constr.* **165**, 105486. <https://doi.org/10.1016/j.autcon.2024.105486> (2024).
9. Yin, Y. & Liu, X. Small target worker detection based on improved YOLOv12 for large construction scenes. *Autom. Constr.* **182**, 106722. <https://doi.org/10.1016/j.autcon.2025.106722> (2026).
10. Fang, Q. et al. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* <https://doi.org/10.1016/j.autcon.2017.09.018> (2018).
11. Y. Gu, S. Xu, Y. Wang, L. Shi, An advanced deep learning approach for safety helmet wearing detection, in: Proc. - 2019 IEEE Int. Congr. Cybermatics 12th IEEE Int. Conf. Internet Things, 15th IEEE Int. Conf. Green Comput. Commun. 12th IEEE Int. Conf. Cyber, Phys. So, 2019. <https://doi.org/10.1109/iThings/GreenCom/CPSCoM/SmartData.2019.00128>.
12. Yang, W., Le Zhou, G., Gu, Z. W., Jiang, X. D. & Lu, Z. M. Safety helmet wearing detection based on an improved YOLOv3 scheme. *Int. J. Innov. Comput. Inf. Control.* <https://doi.org/10.24507/ijic.18.03.973> (2022).
13. D. Yan, L. Wang, Improved YOLOv3 Helmet Detection Algorithm. In 2021 4th Int. Conf. Robot. Control Autom. Eng. RCAE 2021, 2021. <https://doi.org/10.1109/RCAE53607.2021.9638803>.
14. Shen, J. et al. Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning. *Comput. Aided Civ. Infrastruct. Eng.* <https://doi.org/10.1111/mice.12579> (2021).
15. Nath, N. D., Behzadan, A. H. & Paal, S. G. Deep learning for site safety: Real-time detection of personal protective equipment. *Autom. Constr.* <https://doi.org/10.1016/j.autcon.2020.103085> (2020).
16. Wang, Z. et al. Fast personal protective equipment detection for real construction sites using deep learning approaches. *Sensors* <https://doi.org/10.3390/s21103478> (2021).
17. 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.* <https://doi.org/10.1093/jcde/qwad019> (2023).
18. Nguyen, N. T., Tran, Q., Dao, C. H., Nguyen, D. A. & Tran, D. H. Automatic detection of personal protective equipment in construction sites using metaheuristic optimized YOLOv5. *Arab. J. Sci. Eng.* <https://doi.org/10.1007/s13369-023-08700-0> (2024).
19. Park, S., Kim, J., Wang, S. & Kim, J. Effectiveness of image augmentation techniques on non-protective personal equipment detection using YOLOv8. *Appl. Sci.* <https://doi.org/10.3390/app15052631> (2025).
20. Wang, S., Park, S., Kim, J. & Kim, J. Safety helmet monitoring on construction sites using YOLOv10 and advanced transformer architectures with surveillance and body-worn cameras. *J. Constr. Eng. Manag.* <https://doi.org/10.1061/JCEMD4/COENG-16760> (2025).
21. Wang, S. Automated non-PPE detection on construction sites using YOLOv10 and transformer architectures for surveillance and body worn cameras with benchmark datasets. *Sci. Rep.* **15**, 27043. <https://doi.org/10.1038/s41598-025-12468-8> (2025).
22. Wang, S. Domain-adaptive faster R-CNN for non-PPE identification on construction sites from body-worn and general images. *Sci. Rep.* <https://doi.org/10.1038/s41598-026-35148-7> (2026).
23. Hua, W. & Chen, Q. A survey of small object detection based on deep learning in aerial images. *Artif. Intell. Rev.* **58**, 162. <https://doi.org/10.1007/s10462-025-11150-9> (2025).
24. Chao, M. et al. A lightweight small object detection model for UAV images based on deep semantic integration. *Sci. Rep.* **15**, 31888. <https://doi.org/10.1038/s41598-025-16878-6> (2025).
25. He, L. & Wei, H. CBAM-YOLOv5: A promising network model for wear particle recognition. *Wirel. Commun. Mob. Comput.* **2023**, 2520933 (2023).
26. Su, K. et al. N-IoU: Better IoU-based bounding box regression loss for object detection. *Neural Comput. Appl.* **36**, 3049–3063. <https://doi.org/10.1007/s00521-023-09133-4> (2024).
27. Zou, Y. & Liu, S. Small object detection algorithm based on improved YOLOv10 for traffic sign. *Transp. Res. Interdiscip. Perspect.* **32**, 101501. <https://doi.org/10.1016/j.trip.2025.101501> (2025).
28. Yang, B., Bender, G., Le, Q. V., Ngiam, J. Condconv: Conditionally parameterized convolutions for efficient inference. *Adv. Neural Inf. Process. Syst.* **32**. (2019).
29. Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In ICASSP 2023–2023 IEEE Int. Conf. Acoust. Speech Signal Process. 1–5 (IEEE, 2023).
30. Wang, S. Domain adaptation using transformer models for automated detection of exterior cladding materials in street view images. *Sci. Rep.* **16**, 2696. <https://doi.org/10.1038/s41598-025-32524-7> (2025).
31. Zheng, Z. et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **52**, 8574–8586 (2021).
32. Ma, S. & Xu, Y. Mpdious: A loss for efficient and accurate bounding box regression. Preprint at <https://arxiv.org/abs/2307.07662>. (2023).
33. SafetyHelmetWearing-Dataset, SafetyHelmetWearing-Dataset, (2019). <https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset?tab=readme-ov-file> (accessed December 4, 2024).
34. S.H.C.V. Project, Safety Helmet Computer Vision Project, (2024). <https://universe.roboflow.com/digitalimage-114q6/safety-helmet-q3b8o> (accessed December 4, 2024).
35. Hard Hat Workers Dataset, Hard Hat Workers Dataset, (2022). <https://public.roboflow.com/object-detection/hard-hat-workers> (accessed December 4, 2024).
36. Wang, S. Development of an automated transformer-based text analysis framework for monitoring fire door defects in buildings. *Sci. Rep.* **15**, 43910. <https://doi.org/10.1038/s41598-025-27648-9> (2025).
37. Zoph, B. et al. Rethinking pre-training and self-training. *Adv. Neural Inf. Process. Syst.* **33**, 3833–3845 (2020).
38. Li, L.H. et al. Grounded language-image pre-training. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 10965–10975. (2022).
39. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L. Dense contrastive learning for self-supervised visual pre-training. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 3024–3033. (2021).
40. Wang, S., Kim, M., Hae, H., Cao, M. & Kim, J. The development of a rebar-counting model for reinforced concrete columns: Using an unmanned aerial vehicle and deep-learning approach. *J. Constr. Eng. Manag.* **149**, 1–13. <https://doi.org/10.1061/JCEMD4.COEN-13686> (2023).
41. Srinivasu, P. N., Kumari, G. L. A., Narahari, S. C., Ahmed, S. & Alhumam, A. Exploring the impact of hyperparameter and data augmentation in YOLO V10 for accurate bone fracture detection from X-ray images. *Sci. Rep.* **15**, 9828. <https://doi.org/10.1038/s41598-025-93505-4> (2025).
42. Behzadian, A., Muturi, T., Owor, N. J., & Adu-Gyamfi, Y. Optimizing road damage detection with YOLOv10: A resource-efficient approach utilizing augmentation, data sampling, and hyperparameter tuning. In 2024 IEEE Int. Conf. Big Data. 8439–8446. (2024). <https://doi.org/10.1109/BigData62323.2024.10825391>.
43. Bengamra, S., Mzoughi, O., Bigand, A. & Zagrouba, E. A comprehensive survey on object detection in visual art: Taxonomy and challenge. *Multimed. Tools Appl.* **83**, 14637–14670 (2024).
44. Arkin, E., Yadikar, N., Xu, X., Aysa, A. & Ubul, K. A survey: Object detection methods from CNN to transformer. *Multimed. Tools Appl.* **82**, 21353–21383 (2023).

45. Wen, L., Cheng, Y., Fang, Y. & Li, X. A comprehensive survey of oriented object detection in remote sensing images. *Expert Syst. Appl.* **224**, 119960. <https://doi.org/10.1016/j.eswa.2023.119960> (2023).
46. Chen, K. et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. (2019). <http://arxiv.org/abs/1906.07155>.
47. Rath, S. YOLOv8 Ultralytics: State-of-the-Art YOLO Models, <https://Learnopencv.Com/Ultralytics-Yolov8/>. (2023).

Author contributions

Seunghyeon Wang: Data curation; Formal analysis; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing—original draft; Writing—review and editing. Hyunju Lee: Project administration; Resources; Validation; Writing—original draft. Young Yeo: Formal analysis; Funding acquisition; Resources. Juhyung Kim: Funding acquisition; Data curation; Formal analysis; Investigation.

Funding

This research was supported by the National Research Foundation of Korea (NRF), funded by the Korean government (Ministry of Science and ICT) under Grant No: RS202400356697.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

All methods in the present study were carried out in accordance with relevant guidelines and regulations. The data used in this study were obtained from a previously released body-worn camera dataset and three publicly available datasets. In the present study, no new image or video data were collected, and there was no direct interaction with human participants. The original on-site data collection for the body-worn camera dataset was conducted with permission from the participating construction companies, and informed consent was obtained from the participating workers, as reported in the original source publications. Informed consent was obtained from all identifiable participants, or their legal guardians where applicable, for the publication of the relevant images and related information in an online open-access publication, as documented in the original source publications and/or dataset repositories. No participant names, patient names, or direct personal identifiers are included in the manuscript, figures, tables, or supplementary materials. Publicly available secondary datasets were used in accordance with the original sources' stated conditions and applicable licenses/terms of use.

Additional information

Correspondence and requests for materials should be addressed to J.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026