Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

# Performance Comparison of Six YOLO Models for Personal Protective Equipment (PPE) Detection

**Aisyah Widayani[1*], Moh. Rizanto Juliarsyah[2], Shofiyatur Rohmah[3], Herman Bagus Dwicahyo[4]**
[1,3,4] Faculty of Vocational Studies, Universitas Airlangga, Indonesia, [2]Politeknik Negeri Malang, Indonesia
[1*]aisyahwidayani@vokasi.unair.ac.id, [2]mohammad.rizanto@polinema.ac.id, [3]shofiyatur.rohmah@vokasi.unair.ac.id, [4]herman.bagus.dwicahyo@vokasi.unair.ac.id

open access

*Corresponding Author

**Article History** :
Submitted: 31-10-2026
Accepted: 13-11-2026
**Published: 19-11-2026**

**Keywords:**
Object Detection; YOLO Models; Personal Protective Equipment; Localization Performance; Construction Safety.

**Brilliance: Research of Artificial Intelligence** is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

## ABSTRACT

Visual safety monitoring requires both recognition of missing personal protective equipment and accurate localization of the relevant evidence. High image accuracy may obscure poor object detection when a single prediction is sufficient to classify an entire photograph. This study examines that discrepancy through benchmarking of six lightweight detectors on a curated public construction image dataset. A duplicate audit retained 1,402 images, divided nto 1,118 training, 143 validation, and 141 test images. Each detector was trained with three random seeds under a common fine-tuning technique. Checkpoints and image decision thresholds were selected using validation data before test evaluation. Outcomes included image accuracy, mean average precision across overlap thresholds from 0.50 to 0.95, and class-specific localization performance. Across the six detectors, test image accuracy ranged from 98.35% to 99.29%, whereas all-class mean average precision ranged from 26.90% to 27.94%. Mean average precision for the four missing-equipment categories was only 3.90% to 5.61%. In one representative run, all 141 image decisions were correct, but only 13 of 154 missing-equipment instances were correctly localized and classified at the stated object-matching thresholds. These findings show that accurate image screening does not establish reliable worker-directed monitoring. Separate image and object evaluations are therefore necessary, together with expert review of label meaning, person-equipment associations, and site requirements before operational use.

## INTRODUCTION

Visual monitoring of personal protective equipment (PPE) requires more than recognizing an image that contains missing equipment. A detector may flag an uncovered head without locating the relevant person, identifying other absent items, or establishing whether equipment is required. Image screening can prioritize inspection, whereas worker-directed intervention needs localization, association, and worksite context. Evaluation must distinguish these tasks before accuracy is interpreted as effective safety monitoring. This follows the problem-aware metric selection discussed by Reinke et al. (2024) and Maier-Hein et al. (2024), applying their general evaluation principles beyond the original biomedical setting.

Lightweight detectors are attractive because they can be trained on modest hardware. Recent developments include the training and scaling strategies of YOLOv7 (C.-Y. Wang, Bochkovskiy, & Liao, 2023) and the programmable gradient information of YOLOv9 (C.-Y. Wang, Yeh, & Liao, 2025). These changes motivate controlled comparison, but generic benchmark gains do not establish reliable detection of safety-relevant PPE categories.

Two evaluation choices deserve particular attention. First, the unit of analysis may be an object or a whole image. Object detection requires class agreement and spatial overlap, whereas image screening may count a single confident missing-PPE prediction anywhere in a photograph. Second, the meaning of a missing-PPE label depends on the dataset. An annotation such as no_helmet records an assigned visual category; it does not independently establish an occupational requirement, a breach of a site procedure, or the cause of a person's behaviour. Conflating these choices can produce a stronger operational claim than the experiment supports.

The present study examines this issue using a curated derivative of the public Construction-PPE dataset (Dalvi, Singh, Bhingarde, & Chalke, 2025). Six nano-scale detectors are compared such as YOLOv5nu, YOLOv8n, YOLOv10n, YOLO11n, YOLO12n, and YOLO26n. The experiment uses a common fine-tuning technique, preserves the source taxonomy, audits image duplication, locks operating thresholds on validation data, and retains predictions for inspection. Its central question is whether strong image screening performance is accompanied by strong localization of the explicit missing-PPE classes.

The contribution is an evaluation of how metric choice changes the interpretation of the same predictions. Class-specific average precision, object confusion counts, and a fixed-threshold comparison are examined alongside image

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

decisions. A conservative association prototype illustrates the additional evidence needed for person-level interpretation; it is not an independently validated safety classifier.

Three questions guide the analysis: how the six detectors compare under one training technique, whether image accuracy conceals weak missing-equipment detection, and which conclusions remain defensible after accounting for duplication, class imbalance, threshold selection, and missing worksite context. All three planned seeds are reported for every detector. The study assesses visual PPE evidence rather than injury prevention or comprehensive occupational safety performance.

## LITERATURE REVIEW

Recent PPE studies use different datasets and endpoints. Ferdous and Ahsan (2022) developed a YOLO construction detector and introduced CHVG (four colored hardhats, vest, safety glass). Lo, Lin, and Hung (2023) compared YOLO implementations for helmets and vests, while Nazli et al. (2024) developed a YOLOv5 monitoring system. SH17 (Safe Human 17 dataset) contains 8,099 images and 17 industrial categories (Ahmad & Rahimi, 2025); it offers a possible external test after class matching but is not evaluated here.

Architecture and training refinements address difficult categories. Barlybayev et al. (2024) compared YOLOv8 variants; Tong, Li, Bu, Wang, and Yu (2025) and S. Wang, Kim, Yeo, and Kim (2026) examined multi-scale mechanisms based on YOLOv8n and YOLOv10. Utomo, Adi, Sasongko, and Rahman (2025) studied tuning and augmentation for underrepresented classes. Hygiene monitoring (Alashrafi et al., 2025) and extreme construction conditions (Ding & Luo, 2024) introduce different domains. These studies motivate class-specific analysis, but their scores are not direct baselines for the present split.

Person-level interpretation requires additional evidence. Vukicevic et al. (2022) used body regions, Li, Zhao, Zhou, and Zhang (2022) incorporated pose information, and Cheng, Wong, Luo, Wang, and Leung (2022) combined re-identification with PPE classification. Chern, Hyeon, Nguyen, Asari, and Kim (2023) considered worksite context. Object and keypoint detection (Yang, Xiao, & Zhang, 2024) and detection with pose estimation (López, Suárez-Ramírez, Alemán-Flores, & Monzón, 2025) also motivate explicit associations. Construction-PPE lacks independently validated references for all these relationships and requirements.

The six tested implementations span different YOLO designs. YOLOv5u uses an anchor-free head and differs from the original anchor-based YOLOv5 (Ultralytics, 2023b). YOLOv8 and YOLOv11 provide established comparison points (Ultralytics, 2023a, 2024). A. Wang et al. (2024) introduced consistent dual assignments in YOLOv10; Tian, Ye, and Doermann (2025) emphasized attention-based processing in YOLOv12. Jocher et al. (2026) described the end-to-end YOLO26 design and revised regression head in a preprint. These sources explain implementation selection without predicting a ranking on Construction-PPE. RT-DETR provides a recent transformer alternative (Zhao et al., 2024), but was not trained here; neither were YOLOv7 or YOLOv9.

The study complements architecture-focused work by testing whether the evaluation endpoint matches the intended decision. Stored predictions support image screening, object counts, and qualitative inspection. The source vocabulary is retained (Dalvi et al., 2025), while occupational requirements are treated as additional information. The contribution is diagnostic rather than a new architecture or a validated compliance system.

## METHOD

### Research workflow

Figure 1 summarizes the research workflow. Image and annotation checks and a duplicate audit precede controlled fine-tuning of six nano detectors with three seeds each. Validation data determine the checkpoints, image decision thresholds, and final model choice before test evaluation. Saved predictions support separate object-level and image-level analyses, followed by uncertainty estimation, artifact checks, and qualitative interpretation. Worker-equipment association remains an illustrative prototype without a validated worker-level accuracy estimate.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

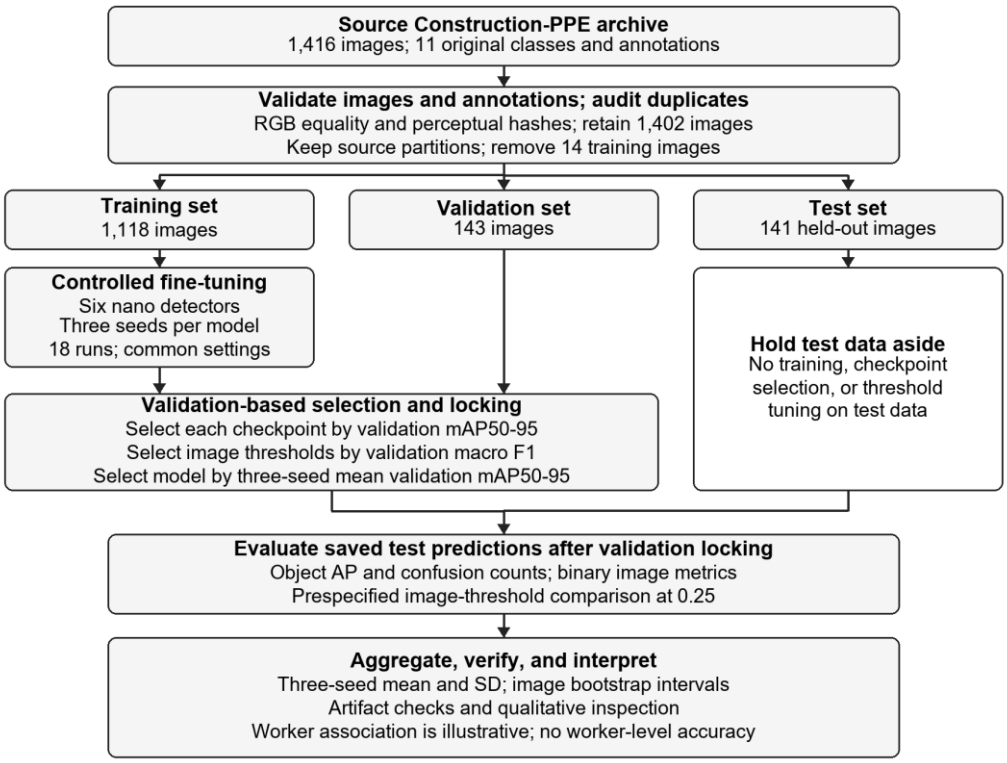



Figure 1. Flowchart of the research methodology. The test split is reserved for evaluation after validation-based selection and threshold locking. Object localization and image screening are evaluated separately; worker-level interpretation is illustrative.

**Dataset provenance and label semantics**

The source archive contains 1,416 images, 1,132 for training, 143 validation, and 141 test images. Its 11 classes are helmet, gloves, vest, boots, goggles, none, Person, no_helmet, no_goggle, no_gloves, and no_boots. Original class identifiers and annotations are retained. The publisher identifies the dataset under AGPL-3.0; the source URL, archive checksum, and license are stored with the local research package. This is a curated derivative of public images, not a newly collected workplace sample.

The binary image target is defined as the presence of at least one source annotation in the four PPE non-use classes. A negative image has no such annotation; it is not labelled fully safe. The class none is excluded from this binary definition because its meaning does not establish compliance. There is no explicit no_vest category. Consequently, failing to detect a vest cannot be converted into a validated missing-vest target using this dataset alone. Every numerical reference to missing PPE in the following sections uses this restricted annotation-based definition.

A pre-test review of training images found absent-helmet, eyewear, and glove annotations in a dining scene. This illustrates why source labels alone cannot establish occupational requirements. Annotations, partitions, and training settings were retained under the predefined protocol. The observation informed interpretation and the association prototype; independent expert review remains necessary to assess annotation relevance and consistency.

**Duplicate audit and retained splits**

Each source image was checked for readability, annotation availability, valid class identifiers, and normalized box coordinates. Duplicate candidates were identified either by equality of decoded RGB pixels or by simultaneous proximity of 64-bit perceptual and difference hashes. Both hash distances had to be at most two bits, and the relative aspect-ratio difference had to remain below three percent. Candidate pairs were joined into connected components. One representative was retained in each component, prioritizing test images, then validation images, then training images. This deterministic rule prevents an identified duplicate from remaining on both sides of a split.

The audit removed 14 training images and identified six cross-split candidate pairs. The retained 1,402 images comprise 1,118 training, 143 validation, and 141 test images. A manifest records source paths, splits, annotations, hashes, grouping, and retention decisions. Counts are reproducible from the manifest, but these checks do not establish the semantic validity of every box.

Verified site and camera identifiers are unavailable. Similar backgrounds, people, and sequences may therefore remain across splits despite the hash audit. The official partition is retained without inferring unsupported provenance

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

groups. Test results describe this source split rather than unseen sites, a limitation consistent with the leakage risks discussed by Kapoor and Narayanan (2023). Table 1 reports the retained counts and binary target prevalence.

Table 1. Images and source PPE non-use class annotation prevalence

| Split | Images | Positive | Negative |
|---|---|---|---|
| Train | 1,118 | 240 | 878 |
| Validation | 143 | 28 | 115 |
| Test | 141 | 26 | 115 |

The four missing-PPE classes contain 400, 337, 442 and 88 training boxes, respectively; the corresponding validation counts are 45, 41, 56 and 4, and the test counts are 40, 33, 58 and 23. These are instance counts, not counts of independent workers. Multiple classes and multiple instances can occur in the same photograph. Positive-image prevalence alone therefore does not describe the support available for learning or evaluating each localized category.

**Models and controlled fine-tuning**

All six models use official COCO-pretrained nano weights, followed by adaptation to the 11 dataset classes. Transfer learning initializes general visual features, while fine-tuning learns the source PPE categories. The tested software is Ultralytics 8.4.178 with PyTorch 2.7.1+cu128(CUDA 12.8), Python 3.12.10, and an NVIDIA GeForce RTX 4050 Laptop GPU with approximately 6 GB of device memory. PyTorch provides the numerical training backend.

The design contains 18 runs: six models with seeds 42, 123, and 2026. All six models are run for one seed before the next seed begins. Each run permits 60 epochs with early-stopping patience of 15. Input size is 512 pixels, physical batch size is eight, and data-loader workers are two. Deterministic settings and automatic mixed precision are enabled. The nominal batch is 64, allowing framework gradient accumulation; eight is the physical batch, not necessarily the effective update batch.

AdamW is used with initial learning rate 0.001, final learning-rate fraction 0.01, weight decay 0.0005, three warm-up epochs, and cosine scheduling. The augmentation settings are fixed across models: mosaic probability 0.5, mosaic closure for the last ten scheduled epochs, translation 0.1, scale 0.5, and horizontal-flip probability 0.5. Rotation, vertical flips, mixup, and copy-paste are disabled. Validation and test images receive no training augmentation. Other implementation defaults are retained in the archived run arguments.

This common technique compares fine-tuning under a fixed budget rather than each architecture's maximum-performance training procedure. The AdamW override differs from the YOLO26 optimizer described by Jocher et al. (2026). Pretraining, heads, assignment, and losses remain architecture-dependent, so results describe complete model and technique combinations. Table 2 summarizes the controls.

Table 2. Common training and evaluation settings

| Setting | Value |
|---|---|
| Models | Six official nano variants |
| Training seeds | 42, 123, 2026 |
| Epochs/patience | 60 maximum / 15 |
| Image/batch | 512 pixels / 8 images |
| Optimizer | AdamW; lr 0.001; decay 0.0005 |
| Schedule | Cosine; warm-up 3; final lr fraction 0.01 |
| Mosaic/close | 0.5 / last 10 scheduled epochs |
| Selection | Validation mAP50-95 |
| Image threshold | Validation macro F1; fixed 0.25 comparison |
| Object matrix | Confidence 0.25; IoU 0.50 |
| Test images | 141 shared across all models |

**Checkpoint and operating-threshold selection**

The best checkpoint in each run is selected using validation mAP50-95, the mean of average precision across classes and intersection over union (IoU) thresholds from 0.50 to 0.95 in steps of 0.05. The test split is not used for epoch selection. The final model choice is defined as the highest mean validation mAP50-95 across the three completed seeds. Seed 42 is fixed as the representative run for learning curves and qualitative inspection. All 18 runs passed the artifact consistency checks before final aggregation.

For image screening, the score is the largest confidence assigned to a prediction in any of the four missing-PPE classes, or zero if none is retained. A positive decision requires a score at least equal to the operating threshold. Epoch monitoring uses 0.25. For held-out analysis, the threshold is selected on validation data from 0.05 to 0.95 in increments of 0.01 by maximizing macro F1. Ties favour higher positive-class recall and then the threshold closest to 0.25. The selected threshold is recorded and locked before test evaluation.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

A prespecified comparison at threshold 0.25 measures operating-point sensitivity using the same saved predictions. Test performance is not used to choose the threshold. A constant-negative classifier provides a prevalence-based reference with zero positive recall.

**Prediction consistency and outcome measures**

Validation and test predictions are stored through the detection validator at confidence 0.001, with suppression overlap threshold 0.7 where applicable. Compatible heads retain the validator's multi-label behaviour; end-to-end heads retain their supported inference path. This avoids discrepancies with a general prediction interface that may use different defaults. The same stream supplies image scores, object matrices, and illustrations, with coordinates mapped to original image dimensions.

The primary detection outcome is test mAP50-95; mAP50 and per-class AP accompany it. A separate descriptive mean of AP50-95 over no_helmet, no_goggle, no_gloves, and no_boots is calculated to expose the behaviour of the missing-PPE subset. This subset summary is an exploratory diagnostic, not a replacement for the prespecified primary endpoint or a basis for retraining the models. Image outcomes comprise accuracy, macro F1, balanced accuracy, and positive-class precision and recall. Standard classification calculations use scikit-learn.

The object confusion matrix uses a confidence threshold of 0.25 and greedy one-to-one matching of candidate pairs in descending intersection-over-union order, accepting overlaps of at least 0.50. Matching is class-agnostic, so an overlapping wrong-class prediction is visible as an off-diagonal entry. Rows represent ground truth, and columns represent predictions. Unmatched annotations and predictions are assigned to the background column and row, respectively. Background-to-background true negatives are undefined for this object-level representation and are displayed as zero. These counts are distinct from the official average-precision calculation.

**Uncertainty and reproducibility checks**

Results are mean and sample standard deviation across three seeds on the same split. Image-metric uncertainty uses 2,000 stratified bootstrap resamples of test images, bootstrap seed 2026, and the 2.5th and 97.5th percentiles. This estimates sampling variation within the observed data, not generalization to unseen sites. No average-precision confidence interval is supplied.

Perfect image predictions yield a degenerate empirical bootstrap interval at 100%, which cannot establish perfect population performance. The three runs reuse the same 141 images and cannot be pooled as 423 independent photographs. Between-seed variation is retained, without statistical-significance or universal-superiority claims.

Artifact checks verify counts and unique image identifiers, validation-test separation, confusion totals, rendered-image counts, finite metric ranges, and checkpoint and dataset-manifest checksums. Training curves use optimizer epochs from the training CSV; a final best-checkpoint callback is excluded from the epoch axis but retained in the event logs. These accounting checks support transparent reporting (Kapoor et al., 2024) and do not replace independent annotation or external validation.

**Conservative worker-level interpretation**

The demonstration associates an equipment box with a person only when its centre lies inside the person box and at least 80 percent of its area is covered by that person. Approximate anatomical position adds constraints for head equipment and boots. An association is withheld if multiple people are plausible matches. Conflicting present and absent labels also cause abstention. The prototype accepts the site's required equipment as an explicit input, with helmet and vest used as an illustrative configuration rather than a universal occupational rule.

Worker-level compliance labels remain disabled until the worksite context and required equipment are explicitly confirmed. After confirmation, a detected required missing item can indicate noncompliance; visible and uniquely associated equipment can support compliance only for the assessed items. Occlusion, conflicting predictions, and missing evidence produce a needs-review status. No worker-level accuracy is reported because the dataset lacks independently reviewed person-equipment links and standard operating procedure (SOP) requirements. A missing-PPE box is model evidence, not a conclusion about negligence.

## RESULT

**Detection and image screening**

Table 3 presents the performance of six YOLO models based on three training seeds. Overall, all models achieved high image-level accuracy, ranging from 98.35% to 99.29%, with macro F1-scores between 97.21% and 98.82%. However, their test mAP50-95 values were considerably lower, ranging from 26.90% to 27.94%. YOLO12n achieved the highest validation and test mAP50-95 values of 29.97% and 27.94%, respectively, making it the best-performing model based on the predefined validation selection criterion. Meanwhile, YOLO11n obtained the highest macro F1-score (98.82%) and positive recall (97.44%). These results indicate that high image-level classification performance does not necessarily correspond to strong object detection and localization capability, highlighting the importance of evaluating both performance levels independently.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

Table 3. Detection and binary image outcomes in percent as mean ± sample SD across three seeds

| Model | Validation mAP50–95 | Test mAP50–95 | Test mAP50 | Image accuracy | Macro F1 | Positive recall |
|---|---|---|---|---|---|---|
| YOLOv5nu | 28.80 ± 0.25 | 27.13 ± 0.16 | 54.12 ± 0.77 | 99.05 ± 1.08 | 98.37 ± 1.89 | 94.87 ± 5.88 |
| YOLOv8n | 28.76 ± 0.37 | 26.90 ± 0.26 | 53.17 ± 1.11 | 98.82 ± 0.82 | 97.97 ± 1.45 | 93.59 ± 4.44 |
| YOLOv10n | 29.29 ± 0.44 | 27.88 ± 0.54 | 55.35 ± 0.30 | 98.35 ± 0.82 | 97.21 ± 1.38 | 93.59 ± 2.22 |
| YOLO11n | 29.73 ± 0.20 | 27.82 ± 0.50 | 55.21 ± 0.67 | 99.29 ± 0.71 | 98.82 ± 1.18 | 97.44 ± 2.22 |
| YOLO12n | 29.97 ± 0.54 | 27.94 ± 0.39 | 55.98 ± 1.35 | 99.29 ± 0.71 | 98.79 ± 1.22 | 96.15 ± 3.85 |
| YOLO26n | 29.04 ± 0.21 | 27.42 ± 0.40 | 54.43 ± 0.97 | 98.82 ± 0.82 | 97.99 ± 1.47 | 96.15 ± 6.66 |

Figure 2 illustrates the substantial difference between image-level accuracy, overall object detection performance, and missing-PPE localization across the six models. While image accuracy remains consistently high at approximately 98-99%, overall mAP50-95 ranges from 26.90% to 27.94%. More notably, the mean AP50-95 for the four missing-PPE categories is considerably lower, ranging from 3.90% to 5.61%. YOLO12n achieves the highest mean localization performance for these categories, whereas YOLOv8n records the lowest. This considerable gap demonstrates that the models can correctly identify images containing missing-PPE annotations without necessarily detecting and accurately localizing the corresponding objects.

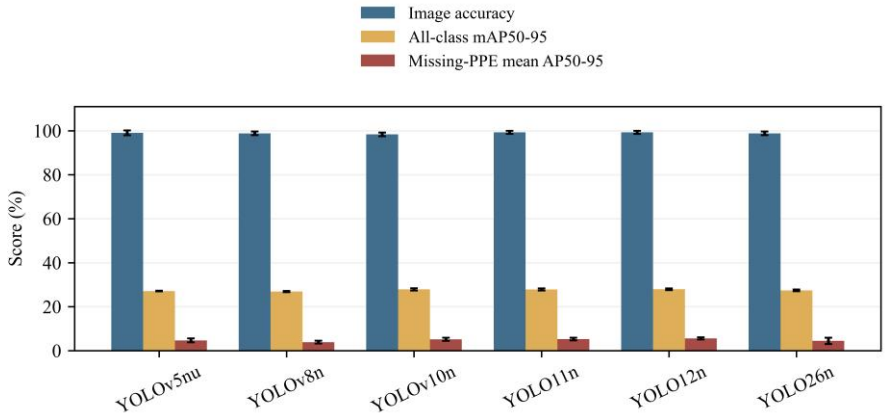



Figure 2. Image screening accuracy, all-class mAP50-95 and mean AP50-95 over the four PPE non-use classes. Bars show three-seed means with SD. The three measures have different evaluation units

Figure 3 shows the validation performance trajectories of the six YOLO models during training using seed 42. The validation mAP50-95 curves generally increase during the early epochs and gradually stabilize toward the later stages, indicating progressive improvements in object localization. In contrast, image screening accuracy shows more noticeable fluctuations but generally reaches higher values during the later training stages. The difference between these two trajectories suggests that image-level classification develops more effectively than object-level localization under the applied training settings. The image accuracy curve uses a fixed confidence threshold of 0.25; therefore, its final value may differ from the test accuracy reported in Table 3, which uses a validation-selected threshold.

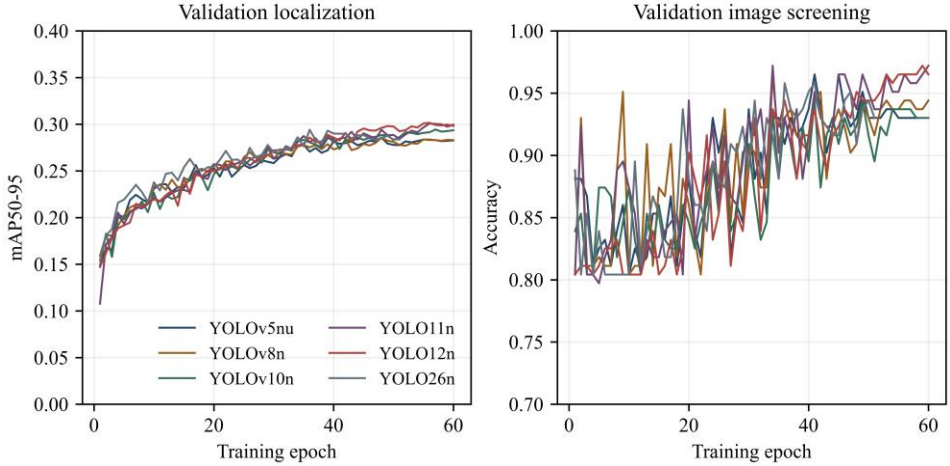



Figure 3. Validation localization and image accuracy during training for seed 42. Image decisions use confidence 0.25. The final best-checkpoint evaluation is not counted as an additional epoch

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

Figure 4 presents the training and validation classification loss trajectories for all six YOLO models using seed 42. Overall, the models demonstrate decreasing loss trends throughout training, with more substantial reductions during the initial epochs followed by gradual stabilization. The training and validation curves generally follow similar patterns, although some differences and fluctuations are observed across architectures. These trends indicate that the models progressively optimize their classification objectives during training. However, the absolute loss values should not be directly compared between architectures because their loss formulations may differ. Furthermore, decreasing classification loss does not necessarily guarantee accurate object localization, which must be assessed separately using detection metrics such as mAP50-95.

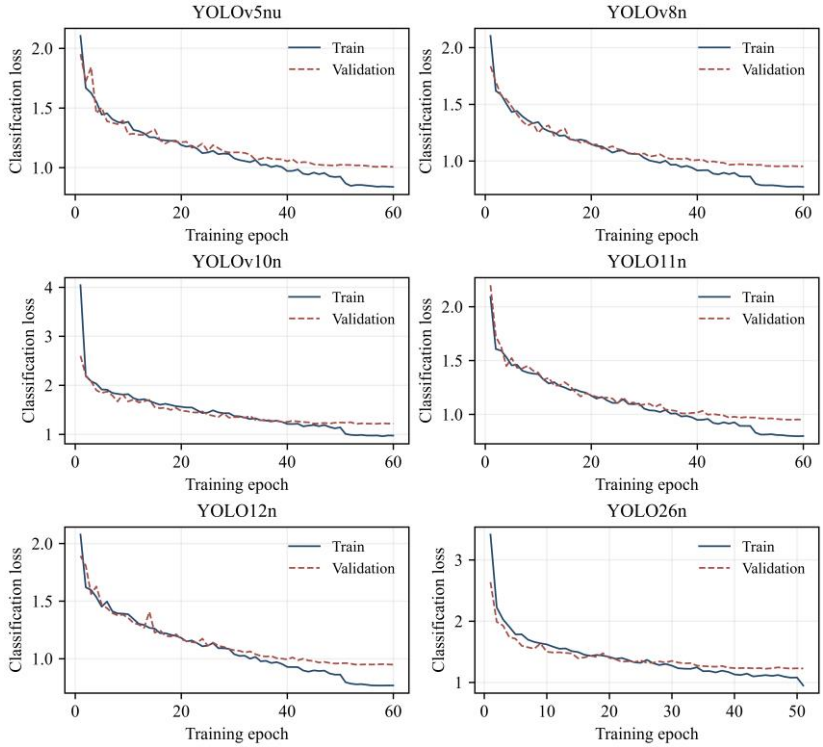



Figure 4. Recorded training and validation classification loss for all six seed-42 runs. Each panel has its own vertical scale; absolute loss magnitudes are not comparable measures of detection quality across architectures

**Missing-equipment localization and error counts**

The descriptive mean AP50-95 over the four PPE non-use classes ranges from 3.90% to 5.61%, substantially below the all-class mean. Table 4 reports all class-level AP values so that this subset summary can be checked. Visible equipment and person categories often account for the stronger part of the overall detection result. The missing-equipment classes therefore remain a material weakness even though an image may contain enough weak evidence to cross its binary screening threshold.

Table 4. Class-specific test AP50-95 in percent as three-seed means

| Class | YOLOv5nu | YOLOv8n | YOLOv10n | YOLO11n | YOLO12n | YOLO26n |
|---|---|---|---|---|---|---|
| helmet | 51.31 | 51.32 | 51.31 | 51.74 | 51.35 | 50.29 |
| gloves | 37.31 | 37.03 | 37.30 | 37.12 | 37.36 | 38.49 |
| vest | 57.51 | 58.12 | 58.13 | 57.38 | 58.23 | 58.45 |
| boots | 38.76 | 38.71 | 39.22 | 39.34 | 38.66 | 39.04 |
| goggles | 31.09 | 30.13 | 33.13 | 32.07 | 31.65 | 31.01 |
| none | 13.27 | 13.69 | 15.01 | 15.19 | 16.92 | 15.19 |
| Person | 50.38 | 51.24 | 51.74 | 52.04 | 50.73 | 51.19 |
| no_helmet | 6.30 | 5.50 | 7.80 | 9.52 | 9.16 | 6.77 |
| no_goggle | 8.32 | 5.11 | 9.00 | 6.88 | 9.22 | 6.71 |
| no_gloves | 3.49 | 4.46 | 3.43 | 3.31 | 3.25 | 3.51 |
| no_boots | 0.73 | 0.53 | 0.56 | 1.45 | 0.80 | 0.96 |

Figure 5 uses YOLOv5nu, the first model in the protocol order, at seed 42 to illustrate the difference between endpoints. Its binary image confusion counts are 115 true negatives, no false positives, no false negatives, and 26 true positives. At the separate object-matrix operating point of confidence 0.25 and overlap 0.50, only 13 of the 154

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

annotated PPE non-use class instances are correctly localized and classified. This example is a diagnostic reading of saved predictions; it does not imply a population-level image accuracy of 100 percent.

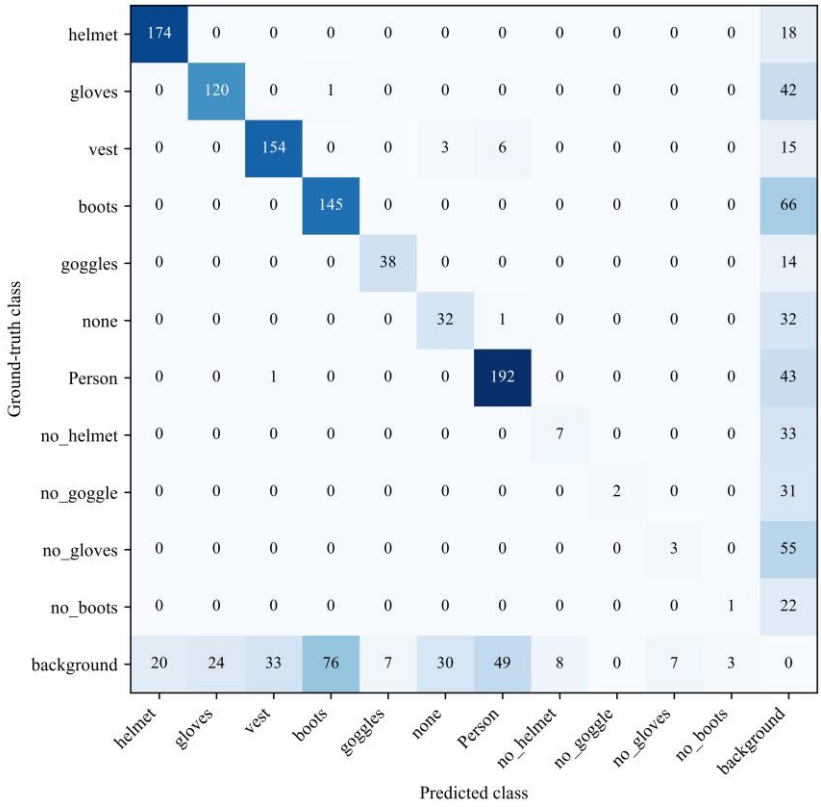



Figure 5. YOLOv5nu seed-42 object confusion counts at confidence 0.25 and IoU 0.50. Rows are ground truth. The last row and column contain unmatched predictions and annotations; background-background is undefined and shown as zero.

**Operating-point sensitivity and qualitative evidence**

Table 5 compares the fixed image threshold with the validation-locked operating point. A constant-negative classifier correctly labels 115 of 141 test images, yielding 81.56% accuracy and zero positive recall. The learned image scores improve on this baseline, but the improvement concerns annotation presence. These results do not estimate worker-level sensitivity because the references do not establish applicable SOP requirements or person-equipment associations.

Table 5. Binary image operating-point comparison in percent as three-seed means

| Model | Accuracy at 0.25 | Recall at 0.25 | Locked accuracy | Locked recall |
|---|---|---|---|---|
| YOLOv5nu | 92.91 | 61.54 | 99.05 | 94.87 |
| YOLOv8n | 94.80 | 71.79 | 98.82 | 93.59 |
| YOLOv10n | 92.91 | 61.54 | 98.35 | 93.59 |
| YOLO11n | 93.38 | 64.10 | 99.29 | 97.44 |
| YOLO12n | 94.56 | 70.51 | 99.29 | 96.15 |
| YOLO26n | 94.33 | 69.23 | 98.82 | 96.15 |

Figure 6 compares all six seed-42 models on test image image1.jpeg, which shows a worker installing solar panels. The illustration was selected after evaluation and did not affect model selection or aggregate metrics. Up to five highest-confidence boxes at or above 0.25 are displayed per model; complete lists are retained in the prediction records. All six models detect the helmet and vest, while four detect gloves above this threshold. This example does not establish full worksite compliance.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

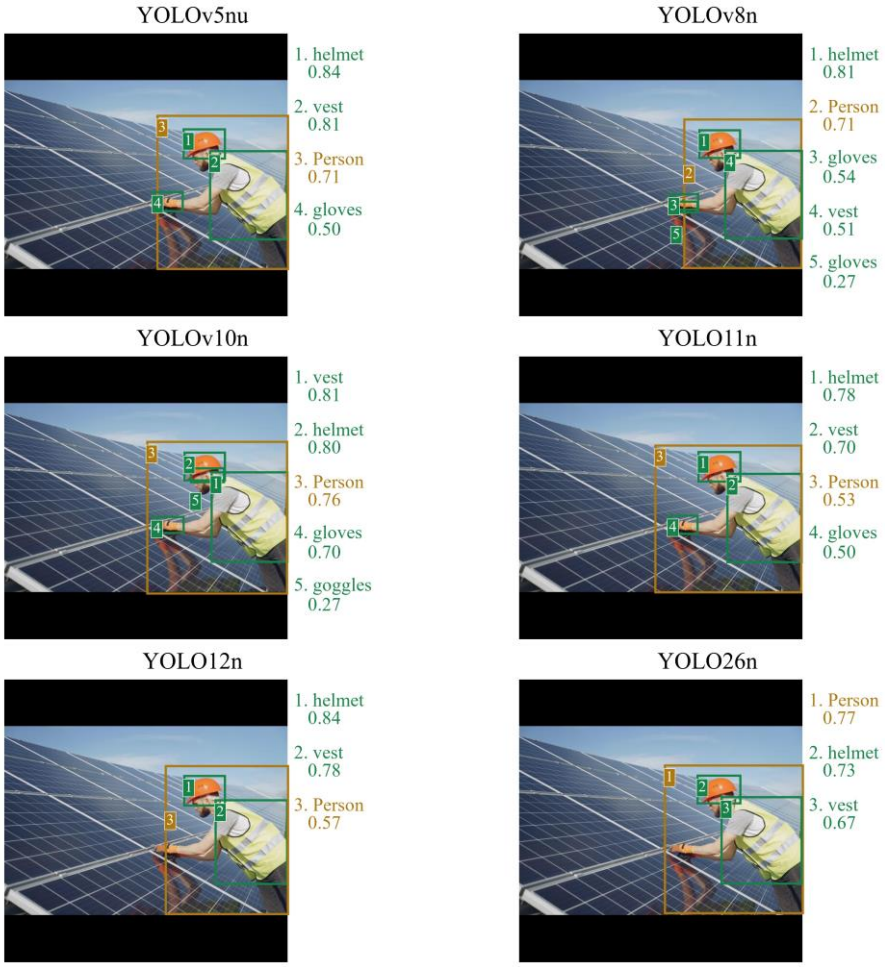



Figure 6. Recorded test predictions from six seed-42 models for a worker wearing PPE during solar-panel installation. Numbered keys give predicted classes and confidence scores. Green denotes PPE classes and ochre the person class. Up to five highest-confidence boxes at confidence ≥ 0.25 are displayed. Visible PPE does not establish full worksite compliance.

## DISCUSSION

### Why the image endpoint is easier

Image screening asks whether any missing-PPE annotation is present. A single positive score can classify an image correctly even when other objects are missed or the predicted box does not overlap the relevant ground truth. Object detection additionally requires spatial and class agreement, and average precision evaluates ranked detections across recall. The numerical gap between these metrics therefore reflects different questions; it is not a calibrated uncertainty estimate.

A correct image decision can therefore coexist with poor recovery of the boxes needed for inspection. This may be acceptable for prioritizing photographs, while worker-specific alerts require stronger evidence. The present analysis extends equipment-recognition work by Ferdous and Ahsan (2022) and Lo et al. (2023) with an explicit check of localized missing-equipment evidence. Different datasets and targets prevent direct numerical ranking against those studies.

The threshold comparison also separates operating-point effects from localization performance. Validation-selected thresholds can recover images with weak missing-PPE scores without changing the predicted boxes or ensuring correct spatial matches. The object matrix uses 0.25, so its counts and the locked image results describe different operating conditions. Low missing-PPE average precision across confidence thresholds shows that the weakness is not confined to the single threshold used in the matrix.

### Class composition and dataset context

Class support is uneven: only four no_boots instances occur in validation, compared with 23 in test. A few difficult objects can therefore change class-specific results substantially. The taxonomy also combines people, visible equipment, and absence categories. Detection metrics weight classes equally, whereas the image target collapses four

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

absence categories into one event. Reporting both the class counts and the aggregation rule is essential to interpreting either result.

Scene context may contribute to the strong image separation. The training-data review found missing-equipment annotations in a dining scene, suggesting that some labels occur outside obvious workplace activity. If those scenes differ systematically from helmet-and-vest scenes, models might distinguish image groups without accurately localizing missing equipment. This remains a hypothesis: no attribution experiment, controlled background intervention, or context-balanced external test was conducted.

Hash-based duplicate removal addresses one identifiable leakage route but does not establish independence between photographs. Images of the same person, scene, or sequence can evade a strict hash rule. The retained split supports reproducible comparison on the source data, while site- or sequence-disjoint evaluation requires additional provenance or independent collection.

**Architecture ranking and computational interpretation**

The ranking applies to this fine-tuning environment. Input size, duration, optimizer, and annotations constrain performance, while heads, losses, and assignment remain architecture-specific. The comparison evaluates complete model and technique combinations, without isolating architectural effects or reproducing each original training procedure.

Loss scales depend on model-specific objectives. Curves assess convergence within a run, while validation detection metrics determine selection and test metrics describe performance. Unusual trajectories require inspection of the corresponding logs.

Validator timings at batch size four exclude acquisition, queuing, rendering, networking, and worker association, and may vary with laptop resource use. They cannot establish camera throughput; deployment requires synchronized end-to-end measurements.

**Implications for an engineering monitoring system**

A monitoring system should retain separate object evidence, image decisions, and person-level interpretations. Boxes, classes, and confidence scores support inspection; identifying a particular worker additionally requires an unambiguous association, visible equipment, and a site-specific requirement. The need for these extra inputs is consistent with body-region analysis (Vukicevic et al., 2022), context-aware assessment (Chern et al., 2023), and detection with pose estimation (López et al., 2025). The present prototype exposes those distinctions but does not validate worker-level compliance.

Abstention avoids forcing a decision when people overlap, body regions are hidden, equipment is held rather than worn, or present-and-absent detections conflict. The geometric rules provide an inspectable starting point, but their thresholds are heuristic. Independently annotated associations and visibility are needed to determine whether abstention improves precision or creates excessive review burden.

For Indonesian occupational health, safety, and environmental management (K3L), the evaluated task covers visual PPE evidence only. Waste handling, chemical spills, machine guarding, evacuation access, and exposure measurement require separate observations and reference outcomes. Any deployment should also define who confirms alerts and how ambiguous evidence enters the inspection process.

**Limitations and next validation steps**

The main limitations are the small test set of 141 images, including only 26 positives; unavailable site identities; and the absence of an independent occupational review of annotation semantics. All trained baselines belong to the YOLO family, and the shared optimizer was not individually tuned for each architecture. The qualitative examples come from the retained test set and provide no independent robustness estimate.

Expert review should address ambiguous and rare categories, occupational context, body-region visibility, required equipment, and person-equipment ownership. Agreement and disagreement resolution need documentation for categorical and spatial annotations. Automated coordinate checks cannot establish occupational meaning.

External evaluation should use sites or acquisition sessions absent from development. SH17 (Ahmad & Rahimi, 2025) is a possible source after class mapping and license review; an independently collected Indonesian site dataset would address the intended local context more directly. Neither extension was evaluated here.

With suitable worker-level references, future tests can compare the prototype with unrestricted matching and forced decisions, measuring review burden at a fixed error tolerance. The present contribution remains the documented mismatch between strong image screening and weak missing-PPE localization.

## CONCLUSION

Test image accuracy ranged from 98.35% to 99.29%, whereas all-class mAP50-95 was 26.90% to 27.94% and missing-PPE mean AP50-95 was 3.90% to 5.61%. These six-model results on the retained Construction-PPE split show

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

**E-ISSN : 2807-9035**
**Volume** 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

that accurate image screening can coexist with poor localization. Image accuracy alone cannot validate worker-directed compliance monitoring.

The completed three-seed validation rule selects YOLO12n. The contribution is a reproducible comparison with explicit endpoints and inspectable errors. Operational use requires expert validation of label meaning, worker-equipment associations, and site requirements, followed by evaluation at unseen locations. Broader K3L claims remain outside the evidence.

## REFERENCES

Ahmad, H. M., & Rahimi, A. (2025). SH17: A dataset for human safety and personal protective equipment detection in manufacturing industry. *Journal of Safety Science and Resilience, 6*(2), 175–185. https://doi.org/10.1016/j.jnlssr.2024.09.002

Alashrafi, L., Badawood, R., Almagrabi, H., Alrige, M., Alharbi, F., & Almatrafi, O. (2025). Benchmarking lightweight YOLO object detectors for real-time hygiene compliance monitoring. *Sensors, 25*(19), 6140. https://doi.org/10.3390/s25196140

Barlybayev, A., Amangeldy, N., Kurmetbek, B., Krak, I., Razakhova, B., Tursynova, N., & Turebayeva, R. (2024). Personal protective equipment detection using YOLOv8 architecture on object detection benchmark datasets: A comparative study. *Cogent Engineering, 11*(1), 2333209. https://doi.org/10.1080/23311916.2024.2333209

Cheng, J. C. P., Wong, P. K.-Y., Luo, H., Wang, M., & Leung, P. H. (2022). Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification. *Automation in Construction, 139*, 104312. https://doi.org/10.1016/j.autcon.2022.104312

Chern, W.-C., Hyeon, J., Nguyen, T. V., Asari, V. K., & Kim, H. (2023). Context-aware safety assessment system for far-field monitoring. *Automation in Construction, 149*, 104779. https://doi.org/10.1016/j.autcon.2023.104779

Dalvi, M., Singh, N., Bhingarde, S., & Chalke, K. (2025). *Construction-PPE: Personal protective equipment detection dataset* (Version 1.0.0) [Data set]. Ultralytics. Retrieved from https://github.com/ultralytics/ultralytics/blob/v8.4.178/docs/en/datasets/detect/construction-ppe.md

Ding, Y., & Luo, X. (2024). Personal protective equipment detection in extreme construction conditions. In Y. Turkan, J. Louis, F. Leite, & S. Ergan (Eds.), *Computing in civil engineering 2023: Resilience, safety, and sustainability* (pp. 672–679). Reston, VA: American Society of Civil Engineers. https://doi.org/10.1061/9780784485248.081

Ferdous, M., & Ahsan, S. M. M. (2022). PPE detector: A YOLO-based architecture to detect personal protective equipment (PPE) for construction sites. *PeerJ Computer Science, 8*, e999. https://doi.org/10.7717/peerj-cs.999

Jocher, G., Qiu, J., Liu, M., Lyu, S., Akyon, F. C., & Kalfaoglu, M. E. (2026). *Ultralytics YOLO26: Unified real-time end-to-end vision models* [Preprint]. https://doi.org/10.48550/arXiv.2606.03748

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., . . . Narayanan, A. (2024). REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances, 10*(18), eadk3452. https://doi.org/10.1126/sciadv.adk3452

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Li, J., Zhao, X., Zhou, G., & Zhang, M. (2022). Standardized use inspection of workers' personal protective equipment based on deep learning. *Safety Science, 150*, 105689. https://doi.org/10.1016/j.ssci.2022.105689

Lo, J.-H., Lin, L.-K., & Hung, C.-C. (2023). Real-time personal protective equipment compliance detection based on deep learning algorithm. *Sustainability, 15*(1), 391. https://doi.org/10.3390/su15010391

López, L., Suárez-Ramírez, J., Alemán-Flores, M., & Monzón, N. (2025). Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation. *Automation in Construction, 176*, 106231. https://doi.org/10.1016/j.autcon.2025.106231

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., . . . Jäger, P. F. (2024). Metrics reloaded: Recommendations for image analysis validation. *Nature Methods, 21*(2), 195–212. https://doi.org/10.1038/s41592-023-02151-z

Nazli, N. A. N. M., Sabri, N., Aminuddin, R., Ibrahim, S., Yusof, S., & Nasir, S. D. N. M. (2024). A real-time system for detecting personal protective equipment compliance using deep learning model YOLOv5. *Procedia Computer Science, 245*, 647–656. https://doi.org/10.1016/j.procs.2024.10.291

Reinke, A., Tizabi, M. D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A. E., . . . Maier-Hein, L. (2024). Understanding metric-related pitfalls in image analysis validation. *Nature Methods, 21*(2), 182–194. https://doi.org/10.1038/s41592-023-02150-0

Tian, Y., Ye, Q., & Doermann, D. (2025). YOLOv12: Attention-centric real-time object detectors. *Advances in Neural Information Processing Systems, 38*, 87151–87175. https://doi.org/10.52202/085713-2627

Tong, B., Li, G., Bu, X., Wang, Y., & Yu, X. (2025). A deep learning-based algorithm for the detection of personal protective equipment. *PLOS ONE, 20*(5), e0322115. https://doi.org/10.1371/journal.pone.0322115

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

**E-ISSN : 2807-9035**
**Volume** 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9939

Ultralytics. (2023a, November 12). *Explore Ultralytics YOLOv8* [Documentation]. Retrieved November 11, 2026, from https://docs.ultralytics.com/models/yolov8/

Ultralytics. (2023b, November 12). *Ultralytics YOLOv5* [Documentation]. Retrieved November 11, 2026, from https://docs.ultralytics.com/models/yolov5/

Ultralytics. (2024, September 30). *Ultralytics YOLO11* [Documentation]. Retrieved November 11, 2026, from https://docs.ultralytics.com/models/yolo11/

Utomo, Z. R., Adi, P. W., Sasongko, P. S., & Rahman, G. (2025). Development and optimization of a construction personal protective equipment (PPE) detection model on YOLOv8 architecture. *Jurnal Masyarakat Informatika, 16*(1), 1–14. https://doi.org/10.14710/jmasif.16.1.71622

Vukicevic, A. M., Djapan, M., Isailovic, V., Milasinovic, D., Savkovic, M., & Milosevic, P. (2022). Generic compliance of industrial PPE by using deep learning techniques. *Safety Science, 148*, 105646. https://doi.org/10.1016/j.ssci.2021.105646

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems, 37*, 107984–108011. https://doi.org/10.52202/079017-3429

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7464–7475). IEEE. https://doi.org/10.1109/cvpr52729.2023.00721

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2025). YOLOv9: Learning what you want to learn using programmable gradient information. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer vision – ECCV 2024* (Lecture Notes in Computer Science, Vol. 15089, pp. 1–21). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-72751-1_1

Wang, S., Kim, H., Yeo, J., & Kim, J. (2026). YOLOv10-based multi-scale variant object detection for multi-category PPE non-compliance monitoring on construction sites. *Scientific Reports, 16*(1), 24591. https://doi.org/10.1038/s41598-026-53052-y

Yang, B., Xiao, H., & Zhang, B. (2024). A unified object and keypoint detection framework for personal protective equipment use. *Developments in the Built Environment, 20*, 100559. https://doi.org/10.1016/j.dibe.2024.100559

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., . . . Chen, J. (2024). DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16965–16974). IEEE. https://doi.org/10.1109/cvpr52733.2024.01605

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.