

SiteShield AI: An Intelligent Video-Based Workplace Safety Monitoring System

Mr. Sangana Koti Reddy¹, Mr. Golla Teja Kiran², Mr. Donta Siddarth³, Mr. Gampala Jaishuk⁴, Mr. Erukali Krishna⁵

¹Professor, Dept. of Computer Science and Engineering, HITAM, Telangana, India

^{2,3,4,5}Student, Dept. of Computer Science and Engineering, HITAM, Telangana, India

Abstract-This paper presents SiteShield AI, an intelligent video-based workplace safety monitoring system for automated detection of workers, personal protective equipment (PPE), and visible PPE non-compliance. The system uses the YOLO11n object detector to identify eight classes: person, helmet, gloves, vest, boots, no_helmet, no_gloves, and no_boots. The dataset contains 1,415 images divided into 1,131 training, 143 validation, and 141 test images, with the independent test set containing 1,101 annotated object instances. Two principal experiments were evaluated: a baseline model trained on the original training data and an augmented model trained with targeted oversampling and image-level augmentation of underrepresented violation classes. The baseline achieved a precision of 0.574, recall of 0.555, mAP@50 of 0.566, and mAP@50:95 of 0.307. The augmented model achieved precision of 0.622, recall of 0.528, mAP@50 of 0.564, and mAP@50:95 of 0.305. The application layer extends detection into worker-level monitoring through PPE association, violation-state stabilization, duplicate suppression, safety-score calculation, live camera visualization, and persistent analysis history. Results show that common PPE classes are substantially easier to detect than rare non-compliance classes, while simple augmentation alone does not solve the rare-class problem. The work demonstrates the feasibility of lightweight YOLO-based safety monitoring while emphasizing the need for richer violation-specific data and human oversight before safety-critical deployment.

Keywords: workplace safety, personal protective equipment, PPE detection, computer vision, YOLO11n, object detection, deep learning, safety compliance, data augmentation, real-time monitoring

1. INTRODUCTION

Workplace safety monitoring is an important application area for computer vision because safety officers may not be able to continuously inspect every worker, camera view, or activity in a large work environment. Personal protective equipment such as helmets, gloves, high-visibility vests, and safety boots is commonly used to reduce exposure to occupational hazards. Manual inspection can identify unsafe

conditions, but it is periodic and dependent on human attention. Automated visual monitoring can complement this process by continuously analyzing images or video frames and identifying objects or conditions that require attention.

Deep-learning object detectors have become practical for this type of application because one-stage detectors can localize multiple objects in a single forward pass. The original YOLO formulation demonstrated the value of treating object detection as a unified regression problem, enabling fast inference compared with multi-stage detection pipelines [1]. Subsequent YOLO generations have been widely applied to construction and workplace safety monitoring [2]–[5].

However, detecting the presence of PPE is not equivalent to detecting non-compliance. Missing-PPE categories can be visually subtle, small, partially occluded, or confused with background. SiteShield AI therefore represents PPE-present and PPE-absent conditions as explicit object classes. This allows the detector to model no_helmet, no_gloves, and no_boots directly rather than inferring non-compliance only from the absence of a positive PPE prediction.

A further challenge is that workplace monitoring is not a single-image classification task. The same worker can appear in many consecutive frames, PPE can be partially hidden, and detector confidence can fluctuate even when the physical situation has not changed. A useful monitoring system must therefore distinguish a persistent safety condition from a temporary detection fluctuation. SiteShield AI addresses this engineering requirement by separating raw detection from worker-level state management and by retaining the strongest evidence while suppressing repeated representations of the same event.

The study is also designed around a reproducible comparison. The independent 141-image test set is kept unchanged while the training procedure is modified through targeted augmentation. This makes it possible to determine whether the additional training data genuinely improves generalization. The resulting experiment is important even though the aggregate mAP does not increase, because it demonstrates that the proposed

augmentation strategy is insufficient by itself for the rarest violation categories.

The research problem is therefore twofold: develop a compact multi-class detector that recognizes workers and PPE conditions, and evaluate whether targeted augmentation can improve the rare violation classes without damaging common-class performance. At the application level, frame-level detections must also be transformed into interpretable worker-level safety information through association, temporal stabilization, duplicate suppression, scoring, and reporting.

- Develop a computer-vision-based workplace safety monitoring system using YOLO11n.
- Detect workers and the presence or absence of major PPE items.
- Train and evaluate the detector on an annotated multi-class PPE dataset.
- Investigate targeted oversampling and image-level augmentation for underrepresented violation classes.
- Associate PPE and violation detections with individual workers for meaningful monitoring.
- Stabilize live violation states, suppress duplicate detections, and calculate a worker-level safety score.

The monitoring problem becomes more difficult when camera distance, worker pose, illumination, occlusion, and background clutter vary. A helmet may occupy a relatively large and distinctive region of a worker image, whereas the absence of gloves or boots can be represented by a very small visual region. Consequently, a system designed for safety monitoring must be evaluated not only on its overall detection score but also on the behavior of individual compliance and non-compliance classes. This study treats the detection task as a measurable research problem rather than assuming that a high aggregate score automatically implies reliable safety decisions.

The proposed work also distinguishes between the machine-learning detector and the application logic surrounding it. The detector is responsible for producing candidate objects and confidence values. The application then converts those candidates into worker-level information, removes repeated detections, maintains stable states across frames, and presents the result through a web interface. This separation makes it possible to evaluate model quality independently from interface behavior and provides a clearer path for future improvements.

The principal contributions of this study are a compact eight-class YOLO11n PPE detector, an experimental comparison between original and targeted-augmented training data, explicit analysis of rare non-compliance classes, and an integrated monitoring layer that converts frame-level detections into worker-level safety information.

The study deliberately reports the negative augmentation result as well as the positive class-level changes so that the conclusions remain consistent with the independent test evidence.

2. LITERATURE SURVEY

Early YOLO research established a fast one-stage approach to object detection by predicting object locations and class probabilities in a unified network [1]. This family of detectors has subsequently been adapted to construction and industrial safety tasks, where speed and the ability to process multiple objects in a frame are important. Studies in workplace and construction environments have reported useful detection of workers, helmets, vests, and other PPE categories [2]–[5].

Related research also identifies important challenges. Real construction imagery contains clutter, varying illumination, scale changes, occlusion, and small objects. PPE non-compliance is particularly difficult because the absence of an item may be represented by a small visual cue rather than a large, distinctive object. Recent work on PPE non-compliance and multi-scale detection further emphasizes the difficulty of detecting small or partially occluded violations [6], while controlled augmentation has been investigated as a way to improve robustness under selected conditions [7].

The present study differs in two practical respects. First, PPE-present and PPE-absent states are explicitly modeled as separate classes. Second, the experiment evaluates targeted oversampling and image-level augmentation against an unchanged independent test set, allowing both positive and negative outcomes to be reported. In addition, the system contributes an application layer that converts object detections into worker-level monitoring information rather than treating bounding boxes as the final output.

An important implication of the reviewed work is that model selection alone does not guarantee field usefulness. A detector can perform well on frequent classes while still missing the events that matter most to safety operators. The present study consequently treats class imbalance, violation recall, and temporal interpretation as first-class evaluation concerns. This provides a more conservative and practically relevant interpretation of the experimental results.

The literature therefore supports the feasibility of YOLO-based PPE monitoring but leaves a practical gap around rare non-compliance classes and the conversion of frame-level predictions into stable operator-facing safety information. SiteShield AI addresses both aspects within a single prototype.

Across these studies, a recurring requirement is realistic training data. Construction and industrial scenes contain

changing backgrounds, partial occlusions, different camera heights, variable worker orientations, and PPE designs that may differ between sites. These factors make direct transfer from a clean dataset to an operational environment uncertain. The literature also shows that object-detection architectures can be effective for PPE-present classes while non-compliance detection remains more difficult because the target is often defined by the absence of an object or by a subtle visual condition.

Another relevant issue is class imbalance. If common classes dominate the training distribution, optimization can favor the classes that contribute the largest number of examples. Oversampling and augmentation are common strategies for reducing this imbalance, but transformed images do not necessarily add new visual information. For safety applications, therefore, augmentation should be evaluated with an untouched test set and with class-wise metrics, not only with a single overall score.

Based on this literature gap, the present study combines detector training with application-level safety logic. The detector provides the visual evidence, while the monitoring layer addresses persistence, duplicate events, worker association, and reporting. This combination is intended to make the research prototype closer to a practical monitoring workflow without claiming that it is a certified autonomous safety system.

3. METHODOLOGY

The methodology follows an end-to-end experimental pipeline consisting of dataset preparation, model training, targeted augmentation, independent evaluation, and application integration. The same 141-image test set is retained for both the baseline and augmented experiments so that the comparison measures model changes rather than changes to the evaluation data.

3.1 Dataset and Annotation

The class distribution is important because the detector is simultaneously learning ordinary objects and rare non-compliance states. The augmented experiment contains only 24 images with no_helmet, 23 images with no_gloves, and 6 images with no_boots in the reported class-wise evaluation. The corresponding instance counts are 40, 58, and 23. These figures show why the no_boots result must be interpreted cautiously: the model has very little visual diversity from which to learn that class, and a small number of difficult examples can strongly influence its measured performance.

Annotation quality is equally important in this setting because a PPE violation is meaningful only when the relevant object or worker condition is correctly localized. The experimental design therefore preserves the annotated

validation and test data rather than generating synthetic evaluation labels. The independent test set is used consistently for both principal models so that class-wise differences can be attributed to the training change rather than to a different test population.

The dataset contains 1,415 images: 1,131 training images, 143 validation images, and 141 test images. The test set contains 1,101 annotated object instances. The eight classes are person, helmet, gloves, vest, boots, no_helmet, no_gloves, and no_boots. The distribution is strongly uneven, with the common PPE classes represented much more frequently than the violation classes; no_boots is especially underrepresented.

The purpose of the targeted strategy is not simply to increase the total number of images. It attempts to expose the model more frequently to visually difficult violation conditions while preserving the original distribution of common classes. Horizontal flipping changes spatial orientation while preserving object identity; brightness, contrast, and HSV changes simulate illumination and appearance variation; blur and noise introduce controlled degradation; and oversampling increases the training frequency of rare violation examples. Because these transformations can still preserve the same underlying visual content, their effectiveness must be established experimentally rather than assumed.

3.2 Preprocessing and Augmentation

Images are represented in YOLO object-detection format and resized to 640×640 pixels during training and evaluation. Targeted augmentation includes horizontal flipping with corresponding bounding-box transformation, brightness and contrast variation, HSV/color variation, slight blur, image noise, and oversampling of images containing no_helmet, no_gloves, and no_boots. Validation and test images remain untouched to prevent evaluation leakage.

3.3 Model and Training

The final baseline checkpoint contains 2,583,712 parameters and approximately 6.4 GFLOPs as reported in the project experiment. Training was performed with workers set to zero and the final evaluation was performed on the independent test set. The use of a lightweight model is aligned with the intended monitoring scenario, where inference efficiency matters, but the research evaluation deliberately prioritizes class-wise correctness rather than claiming a particular deployed frames-per-second rate.

YOLO11n was selected as a lightweight detector suitable for real-time-oriented monitoring. Training used 640×640 inputs, 50 epochs, batch size 4, workers 0, optimizer Auto, initial learning rate

0.01, final learning-rate factor 0.01, momentum 0.937, weight decay

0.0005, warm-up of 3 epochs, mosaic 1.0, horizontal flip probability

0.5, RandAugment, and erasing 0.4. The implementation used Ultralytics 8.4.120, Python 3.14.6, and PyTorch 2.11.0+cu128. The augmented experiment was initialized from the baseline best checkpoint.

Table 1: Principal model training configuration

Parameter	Value
Model	YOLO11n
Input size	640 × 640
Epochs	50
Batch	4
Optimizer	Auto
Initial LR	0.01
Momentum	0.937
Weight decay	0.0005
Warm-up	3 epochs
Mosaic	1.0
Horizontal flip	0.5
Ultralytics	8.4.120

Table 2: Principal training configuration used in the experiments.

Parameter	Configuration
Detector	YOLO11n
Input size	640 × 640
Epochs	50
Batch size	4
Optimizer	Auto
Initial learning rate	0.01
Momentum	0.937
Weight decay	0.0005
Warm-up	3 epochs
Mosaic	1.0
Horizontal flip	0.5
Ultralytics	8.4.120
Python / PyTorch	3.14.6 / 2.11.0+cu128

The comparison is therefore best interpreted as an ablation-style assessment of the targeted training strategy. If the augmented model had improved rare-class recall without sacrificing common-class performance, the result would support the augmentation hypothesis. Instead, the observed changes are mixed: aggregate precision increases, recall decreases, and mAP changes only slightly. Reporting this outcome is important because it prevents the augmentation stage from being presented as successful merely because precision improved.

The experimental design intentionally keeps the evaluation condition fixed. The baseline model is trained from the original training set, while the second model is initialized from the baseline best checkpoint and exposed to a balanced training set with additional transformed examples of rare violation categories. The validation and test sets are not modified. This arrangement reduces the possibility that an apparent gain is caused by changes in the evaluation distribution.

Precision, recall, mAP@50, and mAP@50:95 are used as the principal detector measures. Precision reflects the proportion of reported detections that are correct, while recall reflects the proportion of relevant objects that are recovered. mAP@50 evaluates detection using a 0.50 intersection-over-union threshold, whereas mAP@50:95 averages performance over stricter IoU thresholds. Confusion matrices are also examined because they reveal whether weak classes are being confused with other objects or primarily rejected as background.

4. SYSTEM DESIGN

The architecture intentionally keeps the detector independent from presentation logic. This means a future model can replace YOLO11n without requiring the entire monitoring interface to be redesigned. Conversely, changes to dashboard reporting, history storage, or operator workflows can be made without changing the trained detector. This modularity is useful for a research prototype because it separates model experimentation from application engineering and makes error diagnosis more systematic.

The system is designed as a layered architecture. The visual layer receives an uploaded image or live camera frame. YOLO11n produces person, PPE-present, and PPE-violation detections. The application layer maintains worker tracks, associates nearby PPE detections with workers, stabilizes violation states, suppresses duplicate boxes, calculates a safety score, and stores analysis information for later review. The web layer exposes detection results, live monitoring, statistics, history, analytics, and reports.

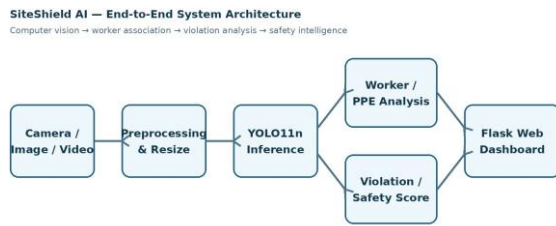


Fig. 1: SiteShield AI system architecture.

4.1 Worker-Level Interpretation

Worker-level interpretation begins with person detections. Stable identifiers are maintained across consecutive frames. PPE and violation detections are associated with the corresponding worker using spatial relationships. This is more informative than simply counting all boxes because one worker may generate multiple PPE detections and multiple frames may contain the same event.

Worker-level interpretation prevents raw box counts from being mistaken for independent safety events. Person detections provide the reference region, nearby PPE detections are associated with that worker, and persistent evidence is retained as a monitoring state. This design lets the detector remain reusable while application rules provide the workplace-specific interpretation.

4.2 Violation Stabilization and Duplicate Suppression

Continuous video produces unstable detections because confidence, bounding-box position, illumination, and occlusion change from frame to frame. SiteShield AI retains a stable violation state and selects the strongest current detection for each violation type. Hysteresis prevents small confidence changes from causing rapid state flicker. A second global suppression layer compares candidate violation boxes using intersection-over-union and center-distance criteria so overlapping boxes representing the same event are reduced to a single displayed violation.

4.3 Safety Compliance Score

The application computes a worker-level safety score rather than treating detector confidence as a safety probability. If workers are present, the score is the percentage of workers currently classified as compliant. The interface maps the score to Excellent (≥ 90), Good (75–89), Attention Required (50–74), and Critical (< 50). This indicator is intended for monitoring and review rather than autonomous safety certification.

5. RESULTS AND EVALUATION

Evaluation uses the same independent test set for both principal experiments. The baseline model achieved precision 0.574, recall 0.555, mAP@50 0.566, and mAP@50:95 0.307. The targeted-augmentation model achieved precision 0.622, recall 0.528, mAP@50 0.564, and mAP@50:95 0.305. Thus, augmentation increased precision but did not improve aggregate mAP or recall.

Table 3: Baseline versus targeted-augmentation evaluation

Metric	Baseline	Augmented
Precision	0.574	0.622
Recall	0.555	0.528
mAP@50	0.566	0.564
mAP@50:95	0.307	0.305

Class-wise evaluation shows a strong difference between common PPE classes and violation classes. In the augmented experiment, person, helmet, gloves, vest, and boots achieved mAP@50 values of 0.839, 0.931, 0.768, 0.901, and 0.742 respectively. The violation classes achieved only 0.215 for no_helmet, 0.108 for no_gloves, and 0.00728 for no_boots.

Table 4: Class-wise performance of the augmented experiment

Class	Images	Inst.	P	R	mAP50	mAP50:95
person	133	236	0.853	0.801	0.839	0.502
helmet	110	192	0.927	0.911	0.931	0.502
gloves	75	163	0.889	0.736	0.768	0.375
vest	111	178	0.863	0.854	0.901	0.568
boots	75	211	0.755	0.678	0.742	0.392
no_helmet	24	40	0.386	0.158	0.215	0.0771
no_gloves	23	58	0.306	0.0862	0.108	0.0241
no_boots	6	23	0	0	0.00728	0.00212

The confusion behavior is consistent with this pattern: normalized recognition is approximately 81% for person, 92% for helmet, 76% for gloves, 84% for vest, and 72% for boots, while no_helmet and no_gloves are much weaker and

no_boots is rarely recognized. The result indicates that the model has learned strong visual representations for common PPE objects but insufficiently reliable representations for rare absence conditions.

The principal aggregate comparison is summarized below. The augmented model increased precision by 0.048 in absolute terms, from 0.574 to 0.622, but recall decreased by 0.027, from 0.555 to 0.528. The change in mAP@50 was -0.002 and the change in mAP@50:95 was -0.002. Therefore, the augmented model can be described as more selective in aggregate, but not as more effective at recovering the complete set of relevant objects.

Table 5: Aggregate independent-test comparison.

Metric	Baseline	Augmented	Change
Precision	0.574	0.622	+0.048
Recall	0.555	0.528	-0.027
mAP@50	0.566	0.564	-0.002
mAP@50:95	0.307	0.305	-0.002

The difference is also visible in recall. Helmet recall reached 0.911 and vest recall 0.854, whereas no_helmet recall was 0.158, no_gloves recall was 0.0862, and no_boots recall was 0. This indicates that many actual violation instances are not recovered by the detector at the reported operating point. In a safety-monitoring application, this limitation is more consequential than a modest increase in precision because missed violations can remain invisible to the operator.

The class-level results provide a more important interpretation for safety monitoring. Helmet reached 0.931 mAP@50 and vest reached 0.901, while the violation classes remained below 0.22 except for no_gloves at 0.108 and no_boots at 0.00728. This means that the detector is substantially better at recognizing the presence of PPE than recognizing the absence conditions that are most directly associated with non-compliance.

Table 6: Functional layers of the integrated SiteShield AI system.

Layer	Main responsibility
Input layer	Uploaded image, video frame, or camera stream
AI layer	YOLO11n detection and confidence generation
Association layer	Worker tracking and PPE-to-worker association
Stabilization layer	Hysteresis, strongest-event selection, duplicate suppression
Analytics layer	Worker counts, violation states,

	compliance score
Storage layer	SQLite analysis/history records
Presentation layer	Flask pages for detection, live view, history, analytics, and reports

A key design principle is that confidence is not treated as a direct safety score. A high-confidence detector output means only that the model is confident in a visual class prediction under its learned representation. The application instead derives a compliance state from worker-level evidence. This prevents the interface from presenting a raw model confidence value as if it were a probability that a worker is safe.

The system also separates detection from persistence. In a live stream, a single missed frame should not immediately erase an otherwise stable violation, and a single weak false positive should not repeatedly create new events. The stabilization and suppression mechanisms are therefore important engineering components of the monitoring prototype, particularly when camera feeds contain natural frame-to-frame variation

6. PRACTICAL APPLICATION

SiteShield AI is designed as a monitoring aid for workplaces in which fixed or temporary cameras observe workers. Its purpose is to reduce repetitive visual screening rather than replace trained safety personnel. An operator can submit an image, inspect an annotated result, or use live monitoring to observe a continuous stream. The system identifies workers and PPE-related detections and converts them into worker-level compliance information.

6.1 Image Analysis and Live Monitoring

For image analysis, the user supplies a workplace image and the application performs preprocessing, YOLO11n inference, confidence filtering, PPE interpretation, and result visualization. For live monitoring, the same detector is repeatedly applied to incoming frames. Worker references and recent violation states can be retained so that the interface represents a continuous scene rather than unrelated frame-by-frame detections.

6.2 Dashboard as an Operational Interface

The dashboard exposes AI engine status, workers detected, PPE compliance, violations, and AI confidence, together with the supported PPE categories. The analysis panel provides a direct entry point for image inspection. This connects the research model to a usable application layer and makes the system easier to demonstrate, debug, and review.

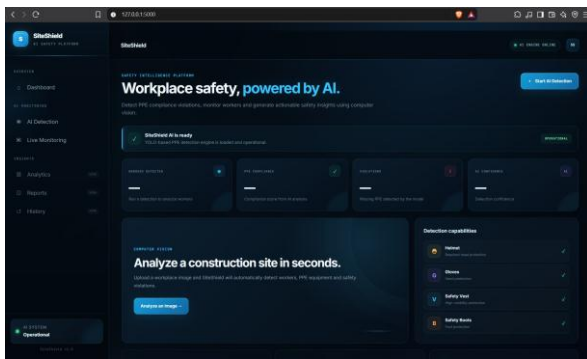


Fig. 2. SiteShield AI dashboard showing engine status, worker/compliance indicators, violations, confidence, and PPE capabilities.

6.3 Practical Safety Use

The platform can act as a first-pass visual screening layer in construction, warehouses, fabrication areas, industrial maintenance, and controlled training environments. A safety team can use it to identify frames or workers that require closer inspection while retaining the original visual evidence for human verification. Its most appropriate role is prioritization and triage, particularly because rare-class recall remains limited.

6.4 History and Reporting

Persistent analysis history provides a practical record of earlier observations. Stored sessions can support internal review, statistics, reporting, model debugging, and repeated demonstrations. The stored record should preserve the distinction between a model-generated observation and a human-confirmed incident so that an automatic alert is not treated as unquestionable ground truth.

6.5 Deployment Scenarios

In a fixed-camera environment, the system can be positioned at entrances, work zones, production areas, or other locations where PPE visibility is reasonably consistent. The camera should provide sufficient worker scale and avoid excessive occlusion. In temporary inspection use, the same software can analyze photographs captured by supervisors. These modes make the platform useful for both continuous observation and periodic review without changing the core detection pipeline.

6.6 Human Review and Decision Support

The application is intentionally suited to human-in-the-loop operation. When a potential violation is displayed, the operator can inspect the underlying frame and determine whether the visual evidence supports the model output. This is particularly important for no_gloves and no_boots, where the measured recall is low and a missed detection

cannot be interpreted as proof of compliance. The dashboard therefore functions as an evidence-prioritization interface rather than an automated enforcement mechanism.

7. SYSTEM WORKFLOW

The workflow is common to uploaded images and live camera frames. It moves from visual acquisition to detection, worker association, violation interpretation, scoring, storage, and presentation. Each stage has a separate responsibility, making the system easier to diagnose and extend.



Fig. 3. SiteShield AI processing workflow from input acquisition through detection, worker association, safety analysis, storage, and presentation.

7.1 Input, Detection, and Tracking

The input stage receives an image or frame and prepares it for the detector. YOLO11n produces bounding boxes, class labels, and confidence values. In continuous monitoring, worker references are maintained across frames so that the same person is not counted as a new worker every time.

7.2 Association, Decision, and Persistence

Person detections act as reference entities for nearby PPE and violation detections. Spatial relationships and confidence information are used to associate relevant objects with the most plausible worker. The application then determines whether a worker has an active violation state. Duplicate suppression and state stabilization reduce repeated representations of the same event, after which the result can be scored, stored, and presented. The separation of layers also clarifies failure analysis. A wrong class is primarily a detector error; a correct PPE box assigned to the wrong worker is an association error; repeated copies of one event are an event-management error. This allows future improvements to target the correct layer instead of assuming every problem requires retraining the detector.

7.3 Stage-by-Stage Processing

The first stages preserve the visual evidence; the middle stages transform detections into worker-level states; and the final stages turn those states into operational records. This

progression prevents the dashboard from being directly coupled to raw neural-network outputs. It also makes debugging more systematic because each stage can be inspected separately when an unexpected event appears.

For an uploaded image, the processing cycle is finite: input is read, detections are produced, worker associations are calculated, and the result is presented. In live video, the same cycle repeats continuously and state is carried between frames. Consequently, live monitoring requires additional attention to temporal consistency, duplicate suppression, and event persistence. These application mechanisms improve usability but do not change the underlying detector metrics.

8. RESULTS AND ANALYTICAL FRAMEWORK

The analytical framework combines aggregate precision and recall with mAP@50, mAP@50:95, class-wise performance, confusion matrices, and application-level event interpretation. Both principal models use the same untouched 141-image test split.

8.1 Aggregate and Class-Level Findings

Train-9 achieves precision 0.574, recall 0.555, mAP@50 0.566, and mAP@50:95 0.307. Train-10-3 increases precision to 0.622, but recall decreases to 0.528 and mAP@50 and mAP@50:95 become 0.564 and 0.305. The 0.048 precision gain is therefore accompanied by a 0.027 recall reduction.

Table 4. Aggregate independent-test comparison.

Metric	Baseline	Augmented	Change
Precision	0.574	0.622	+0.048
Recall	0.555	0.528	-0.027
mAP@50	0.566	0.564	-0.002
mAP@50:95	0.307	0.305	-0.002

The final class-wise results show mAP@50 of 0.839 for person, 0.931 for helmet, 0.768 for gloves, 0.901 for vest, and 0.742 for boots. The violation classes reach only 0.215 for no_helmet, 0.108 for no_gloves, and 0.00728 for no_boots. The detector is therefore substantially stronger for PPE-present recognition than for explicit absence conditions.

8.2 Confusion-Matrix Analysis

Approximately 81% of person, 92% of helmet, 76% of gloves, 84% of vest, and 72% of boots instances are correctly assigned. The corresponding values for no_helmet, no_gloves, and no_boots are about 27%, 16%, and 3%. A single overall mAP value would therefore conceal the most important safety-related weakness.

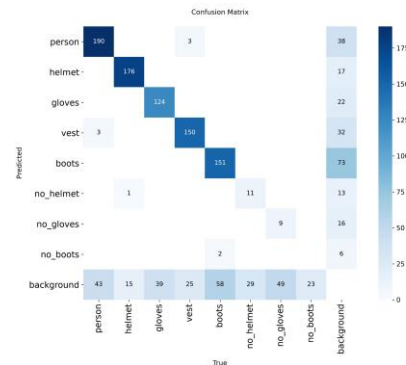


Fig. 4. Test-set confusion matrix from the project evaluation.

8.5 Baseline versus Augmented Interpretation

At the class level, person improves from mAP@50 0.832 to 0.839, helmet from 0.923 to 0.931, gloves from 0.760 to 0.768, and vest from 0.898 to 0.901. Boots decreases from 0.764 to 0.742, no_helmet from 0.236 to 0.215, no_gloves changes from 0.107 to 0.108, and no_boots decreases from 0.0095 to 0.00728. The overall change from 0.566 to 0.564 shows that the small gains in common classes do not compensate for the weak violation categories.

8.6 What the Metrics Mean for Safety Monitoring

The precision increase suggests that the augmented model is more selective at the evaluated operating point, but the lower recall means fewer relevant instances are recovered. For a general object-detection application this may be an acceptable trade-off in some settings; for PPE compliance monitoring it requires caution because missed violations are operationally important. The class-wise results therefore provide a more appropriate basis for deployment decisions than the aggregate precision increase alone.

8.3 Background-Dominant Violation Errors

The normalized matrix shows background proportions of about 0.72 for no_helmet, 0.84 for no_gloves, and 1.00 for no_boots. These values agree with recalls of 0.158, 0.0862, and 0.0. The detector frequently fails to localize the visual evidence required to declare a violation, indicating a representation problem rather than a simple display issue.

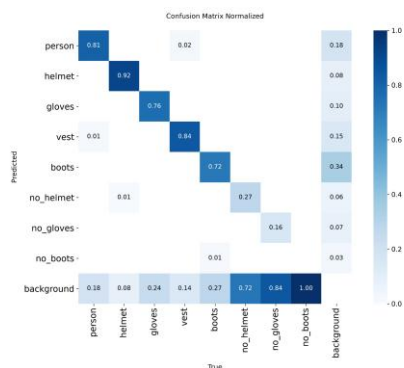


Fig. 5. Normalized confusion matrix showing background classification for rare violation classes.

8.4 Application-Level Analytical Framework

Raw boxes are not treated as independent safety incidents. Worker-level states prevent one worker remaining in view across many frames from artificially multiplying the violation count. Stabilization, strongest-event selection, and duplicate suppression provide a cleaner event representation. Historical sessions can then be compared over time, creating four evaluation levels: detector metrics, class-wise errors, worker-level events, and historical dashboard statistics.

Application logic can improve interpretability, but it cannot recover a violation that the detector never observes. Future evaluation should therefore report event precision, event recall, duplicate-alert rate, persistence duration, and operator confirmation in addition to model-level metrics.

8.5 Error Sources Across the Pipeline

The measured errors can be separated into visual, model, association, and temporal causes. Visual errors arise when the relevant body region is too small, hidden, blurred, or poorly illuminated. Model errors occur when the detector assigns the wrong class or background. Association errors occur when a correct PPE detection is connected to the wrong worker.

8.6 Analytical Implication

The strongest evidence for the current system is therefore the combination of common-class performance and an operational workflow, not a claim of universal PPE compliance accuracy. The model provides useful visual observations, while the application organizes those observations for review. A future evaluation should preserve the same independent-test discipline while adding event-level measures that directly quantify whether the operator receives the correct worker-specific safety information.

9. DISCUSSION AND FUTURE ENHANCEMENT

The detector demonstrates useful performance for several common PPE classes, while the application layer provides a practical path from model inference to worker-level monitoring. At the same time, the rare violation classes remain too weak for autonomous safety decisions. This asymmetry is the central finding of the study and should remain visible when interpreting the dashboard or any future deployment.

9.1 Detection Reliability

Helmet, vest, person, gloves, and boots are comparatively strong because their visual patterns are represented by more examples and are often larger or more distinctive. Missing-PPE categories are different: a missing glove or boot may correspond to a very small region, while the absence itself is expressed through contextual evidence. Occlusion, body pose, motion blur, worker distance, and background clutter further reduce the available signal. For safety monitoring, a false positive can lead to an unnecessary inspection, but a false negative can allow a genuine unsafe condition to remain unnoticed. The decrease in recall from 0.555 to 0.528 therefore matters even though precision improves. The current evidence supports monitoring and triage, not autonomous certification.

9.2 Why Targeted Augmentation Was Insufficient

The targeted experiment was designed to increase representation of no_helmet, no_gloves, and no_boots. Nevertheless, no_helmet mAP@50 decreases from 0.236 to 0.215, no_gloves changes only from 0.107 to 0.108, and no_boots decreases from 0.0095 to 0.00728. Transformed copies of limited examples do not necessarily provide the diversity needed to learn robust visual features.

Keeping the test set untouched is therefore essential. The independent result shows that the augmentation hypothesis is not supported at the aggregate level. This negative result identifies a limitation of simple image transformation and establishes a stronger direction for future data collection.

9.3 Data-Centric Enhancement

The first priority should be substantially more real-world violation examples. New no_helmet, no_gloves, and no_boots images should vary camera distance, viewing angle, worker pose, occlusion, illumination, motion, PPE style, and background. The no_boots class is especially important because its reported evaluation contains only six images and 23 instances, limiting the stability of generalization claims.

9.3.1 Data Diversity Requirements

Additional data should be collected from multiple camera heights and distances and should include workers partially hidden by tools, machinery, materials, or other workers. Images should also represent different PPE colors, designs, and wear conditions. Such diversity is more valuable than repeatedly transforming a small set of nearly identical scenes because the detector must learn the underlying visual variation of real workplaces.

9.3.2 Hard-Negative Mining

Hard negatives should be selected from regions that repeatedly produce false violation detections. Examples may include hands, footwear-like background objects, shadows, machinery edges, and cluttered floor regions. Adding these difficult backgrounds can teach the detector which visual patterns should remain background and can directly target the background-dominant confusion seen in the current experiment.

9.4 Temporal and Application Enhancement

Temporal reasoning is a natural extension for live monitoring. A one-frame violation may be a transient detector error, whereas a condition persisting across several frames provides stronger evidence. A future event-confirmation layer can require consistent worker-level evidence before generating a high-priority alert. Confidence thresholds can also be calibrated separately for positive PPE and violation classes.

The application can record timestamps, worker references, confidence values, evidence frames, event duration, and human confirmation. This auditable event history would support error analysis and threshold calibration and would allow future research to evaluate the monitoring system at the event level rather than only at the object-detection level.

9.5 Field Validation and Deployment

The current experiments do not establish cross-site generalization or end-to-end deployed frame rate. Validation should therefore cover different cameras, lighting conditions, site layouts, worker distances, PPE designs, and occlusion patterns. A staged deployment can begin with shadow monitoring, where predictions are collected without automatically triggering safety actions, followed by review and threshold calibration by safety personnel.

9.6 Research Roadmap

- Collect substantially more violation-specific data before increasing model complexity.
- Use hard-negative mining and class-aware learning to target dominant failure modes.

- Introduce temporal confirmation and worker-level event metrics for live monitoring.
- Calibrate thresholds independently for compliance and violation categories.
- Validate across sites and cameras while retaining human verification.

9.7 Model and Loss-Function Alternatives

Once the dataset is expanded, alternative learning strategies can be evaluated more meaningfully. Class-balanced sampling can reduce the dominance of common classes, while focal or related losses can place greater emphasis on difficult examples. Larger YOLO variants may improve small-object representation, but increasing model size should be considered only after the data limitation is addressed. Each change should be evaluated independently on the same untouched test protocol.

9.8 Operational Validation

Before deployment, a shadow-mode trial can compare system alerts with decisions made by trained safety personnel. Disagreements can be categorized as detector errors, association errors, or operator disagreements. This process can produce a site-specific calibration set and can establish whether the system reduces review workload without introducing unacceptable false-negative behavior.

9.9 Overall Future Direction

The long-term direction is a human-in-the-loop safety intelligence platform in which computer vision performs continuous screening, temporal reasoning confirms persistent conditions, and human reviewers validate consequential events. This approach preserves the efficiency benefits of automation while acknowledging the uncertainty shown by the current rare-class evaluation.

10. CONCLUSION

This paper presented SiteShield AI, an intelligent video-based workplace safety monitoring system combining YOLO11n with worker-level PPE interpretation and a Flask-based dashboard. The detector recognizes person, helmet, gloves, vest, boots, no_helmet, no_gloves, and no_boots, while the application extends those detections into worker association, violation-state management, safety-oriented scoring, visualization, and historical analysis. The dataset contains 1,415 images divided into 1,131 training, 143 validation, and 141 test images. The independent test split contains 1,101 annotated instances and remains unchanged across the baseline and targeted-augmentation experiments. Train-9 achieves precision 0.574, recall 0.555, mAP@50 0.566, and mAP@50:95 0.307. Train-10-3 increases precision to 0.622 but reduces recall to 0.528, while mAP@50 and mAP@50:95 become 0.564 and 0.305.

The results do not support a claim that targeted image augmentation substantially improves overall detection performance. Helmet, vest, person, gloves, and boots remain comparatively strong, whereas no_helmet, no_gloves, and especially no_boots remain weak. The confusion matrices show frequent background classification for the violation categories. The no_boots class reaches only 0.00728 mAP@50 with zero reported precision and recall at the evaluated operating point.

The project demonstrates a complete application path from visual input to operator-facing information: input acquisition, preprocessing, YOLO inference, worker association, violation interpretation, stabilization, duplicate suppression, safety analytics, storage, and dashboard presentation. This layered design permits detector improvements to be evaluated independently from dashboard logic, history, or reporting.

SiteShield AI should therefore be positioned as a monitoring and decision-support aid rather than an autonomous safety authority. The measured rare-class recall is insufficient, and the experiments do not establish cross-site generalization. Human verification, threshold calibration, temporal confirmation, and field validation remain necessary before safety-critical deployment.

The main conclusion is that compact YOLO-based PPE monitoring is technically feasible, but reliable non-compliance detection depends strongly on the quality, diversity, and representation of violation-specific data. Future work should emphasize richer real-world violation examples, hard-negative mining, class-aware training, temporal reasoning, calibrated thresholds, and cross-site evaluation.

10.1 Principal Findings

Three findings summarize the study. First, lightweight YOLO11n detection is feasible for several common PPE categories. Second, targeted augmentation increases aggregate precision but does not improve aggregate mAP or solve rare-class detection. Third, the application layer adds practical value by organizing detections into worker-level states, live monitoring, scoring, history, and dashboard presentation. These findings define both the current capability and the next research priorities.

10.2 Research Significance

The significance of the project is therefore not limited to the numerical score of one model. The work establishes a complete experimental and software framework in which data preparation, model training, class-wise evaluation, application interpretation, and human review can be studied together. The same framework can be reused when

richer data or alternative models become available, making future improvements measurable rather than anecdotal.

10.3 Final Engineering Perspective

A reliable safety-monitoring platform should be judged at both the detector level and the operational event level. A strong detector can still produce an unusable interface if worker association is poor, while a well-designed interface cannot compensate for a detector that misses important violations. SiteShield AI separates these responsibilities, providing a clear foundation for future validation and incremental improvement.

10.4 Integration of Model and Application

The project is significant because the detector and the application are evaluated as separate but connected layers. Model metrics quantify visual recognition, while the dashboard and workflow determine how those observations are interpreted and reviewed. This separation makes future experimentation clearer: a new detector can be evaluated without redesigning the complete interface, and application improvements can be tested without incorrectly attributing their effect to model accuracy.

10.5 Deployment Boundary

The present evidence supports research and assisted-monitoring use, not autonomous safety certification. Before deployment, the model should be tested on representative site imagery, alert thresholds should be calibrated, and human reviewers should confirm consequential events. The system should also maintain an auditable record of model output and operator decisions so that false positives and false negatives can be investigated systematically.

REFERENCES

- [1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," arXiv:1506.02640, 2015.
- [2] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: Real-time detection of personal protective equipment," *Automation in Construction*, vol. 112, 103085, 2020.
- [3] W. Fang et al., "Fast Personal Protective Equipment Detection for RealConstruction Sites Using Deep Learning Approaches," *Sensors*, vol. 21, no. 10, 3478, 2021.
- [4] "PPE detector: a YOLO-based architecture to detect personal protective equipment(PPE) for construction sites," *Scientific Reports / PMC*, 2022.
- [5] "Real-Time Personal Protective Equipment Compliance Detection Based on DeepLearning Algorithm," *Sustainability*, vol. 15, no. 1, 391, 2023.

- [6] S. Wang, H. Kim, J. Yeo, and J. Kim, "YOLOv10-based multi-scale variant object detection for multi-category PPE non-compliance monitoring on construction sites," *Scientific Reports*, vol. 16, 24591, 2026.
- [7] Ö. Akar, H. Selim, O. Er, and D. Kılınç, "AI-Driven Detection of Personal Protective Equipment Usage in Occupational Settings: Enhancing Workplace Safety," 2026.
- [8] G. Jocher and J. Qiu, "Ultralytics YOLO11," Ultralytics software, version 11.0.0, 2024.

The references correspond to the research sources used in the project paper; no separate appendix is included in this version.