

Real-Time Personal Protective Equipment (PPE) Compliance Monitoring System: Comparative Model Evolution and Violation-Centric Design

Trung Do Viet
 Computer Vision_E1402
 Ho Chi Minh City, Viet Nam
 doviettrung1802@gmail.com

Abstract—This report details the design, implementation, and multi-version evolution of a safety-critical Computer Vision (CV) system for automated monitoring of Personal Protective Equipment (PPE) compliance in industrial environments. Model Version 1 (V1) established a detection baseline, but Model Version 2 (V2) introduced the necessary Violation-Centric architecture (detecting NO-PPE). The final iteration, Model Version 3 (V3), addressed critical data deficiencies by increasing the training resolution to 1024×1024 and expanding small-object samples (Gloves). V3 successfully integrated the violation logic with reliable detection across all required classes, resulting in a final model with $\text{mAP50} = 95 = 0.618$ and robust Gloves Recall of 0.828, proving that a strategic data and architecture approach is essential for high-integrity compliance monitoring.

Index Terms—Object Detection, YOLOv9, ByteTrack, Personal Protective Equipment (PPE), Violation-Centric, Safety Compliance, Deep Learning.

I. INTRODUCTION

THE implementation of automated safety monitoring systems is vital for improving industrial site safety and ensuring regulatory compliance. Traditional methods rely on manual supervision, which is prone to human error, inconsistency, and limited coverage. This project develops a robust, deep learning-based system capable of real-time detection and verification of required PPE presence.

This report documents the evolution of the system across three major iterations (V1, V2, and V3), focusing on the architectural changes implemented to address the shortcomings of simple object detection and achieve production-ready compliance output.

II. MODEL VERSION 1 (V1): BASELINE AND LIMITATIONS

A. V1 Architecture and Training

Model V1 employed the **YOLOv9-Compact (YOLOv9-C)** architecture [1] (358 layers, 25.5M parameters). It was trained on a preliminary six-class dataset (Table I) using 800×800 image resolution and a batch size of 4.

B. V1 Conclusion and Deficiencies

The core issue with Model V1 was the **Logic Deficiency**. Compliance was inferred based on the **Presence** of required items. This unreliable logic led to the abandonment of the positive detection approach.

TABLE I
 INSTANCE COUNT FOR MODEL V1 CLASSES

Class Name	Instance Count
Person	1,415
Safety_boots	1,928
Gloves	1,920
Vest	1,514
Hard_hat	1,227
Mask	1,128

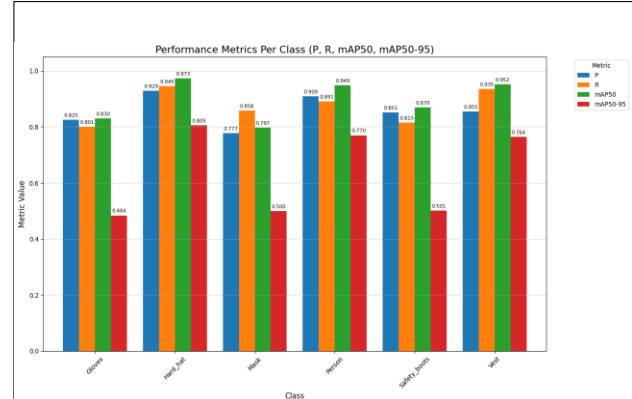


Fig. 1. Performance Metrics per Class (V1). Strong performance for Hard_hat ($\text{mAP50} = 0.973$) but low performance for Mask and Gloves ($\text{mAP50-95} < 0.50$).

III. MODEL VERSION 2 (V2): VIOLATION-CENTRIC ARCHITECTURE

Model V2 addressed the V1 limitations by fundamentally altering both the dataset and the post-processing pipeline.

A. Dataset and Class Expansion

V2 expanded the dataset to include explicit **Negative Classes** ('NO-Hardhat', 'NO-Mask', 'NO-Safety Vest'). The primary goal was to enable direct detection of non-compliance.

The V2 model was trained for 60 epochs using 800×800 images. A known limitation was the severe data sparsity for Gloves ($N = 89$), which resulted in catastrophic Gloves failure (Recall = 0.000).

B. V2 Technical Advancements

- **Violation-Centric Logic:** A person is marked as VIOLATION **only if** the model directly detects the presence of a negative class within the person's bounding box ($\text{IoU} \geq 0.05$)

TABLE II
CLASS INSTANCE COUNT FOR MODEL V2

Class Name	Instance Count	Role
Person	11,013	Region of Interest
Hard_hat	4,609	Compliant Item
NO-Hardhat	2,317	Violation Trigger
Mask	1,651	Compliant Item
NO-Mask	3,097	Violation Trigger
NO-Safety Vest	3,962	Violation Trigger
Gloves	89	Critical Data Flaw

- ByteTrack Integration:** The output pipeline was upgraded using the **ByteTrack** algorithm [3] to assign a persistent tracking ID (e.g., ID: 13) to each worker for robust logging and auditing.

C. V2 Performance Summary

V2 achieved an overall mAP50 – 95 of 0.554, a structural success because it relied on the high accuracy of the violation-triggering classes: NO-Hardhat (mAP50 = 0.817) and NO-Mask (mAP50 = 0.771).

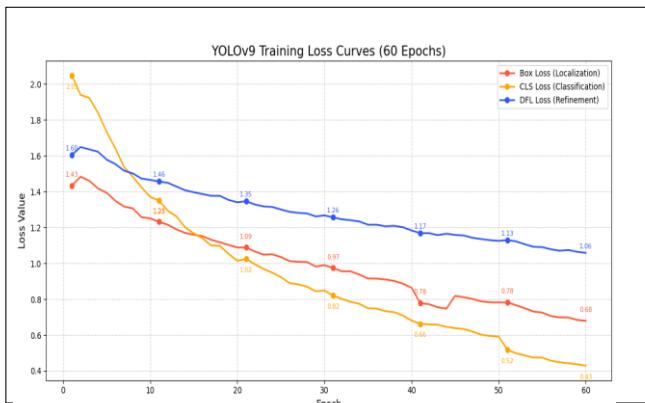


Fig. 2. Training Loss Curves for Model V2 (60 Epochs). All losses showed stable convergence, confirming model health.

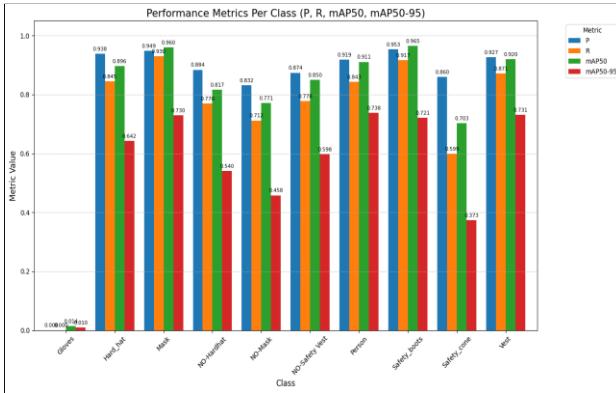


Fig. 3. Performance Metrics per Class (V2). Note the high metrics for violation classes (NO-Hardhat, NO-Mask) which justify the V2 logic.

IV. MODEL VERSION 3 (V3): SMALL OBJECT REMEDIATION

Model V3 was designed solely to fix the data dependency

A. V3 Data and Training Strategy

The V3 strategy prioritized detail acquisition for small objects:

- Data:** Total Gloves instance count was increased from $N = 89$ to $N = 973$.
- Training Resolution:** The image input size ('imgsize') was increased from 800×800 to 1024×1024 to allow the model to learn fine-grain features from the small objects.
- The dataset was constructed by merging custom-collected imager (from construction and factory sites, including those in Vietnam country) with public domain datasets sourced from Roboflow and Kaggle. Initial data preparation involved manual annotation using Roboflow. The combined dataset was then augmented using standard techniques such as random horizontal flipping and rotation to increase sample diversity and generalizability.
- Configuration:** Trained for 60 epochs with YOLOv9-C and batch = 4.

B. V3 Performance Summary

The targeted efforts in V3 resulted in a significant leap in performance and stability across all small object classes (Table III).

TABLE III
KEY METRIC COMPARISON: V2 VS. V3 FOR TARGET CLASSES

Class	V2 R (800px)	V3 R (1024px)	V3 mAP50 (1024px)	V3 mAP50 – 95 (1024px)
Gloves	0.000	0.817	0.858	0.611
Mask	0.881	0.881	0.938	0.675
Safety_boots	0.941	0.941	0.945	0.753
Total	0.727	0.802	0.868	0.618

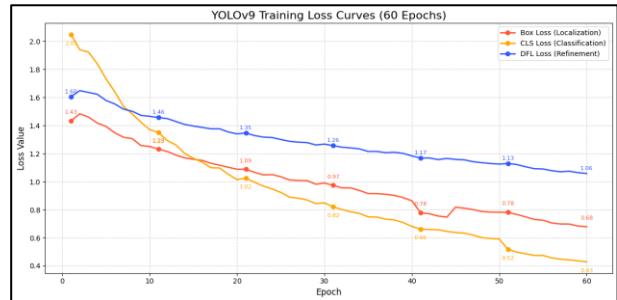


Fig. 4. Training Loss Curves for model V3.

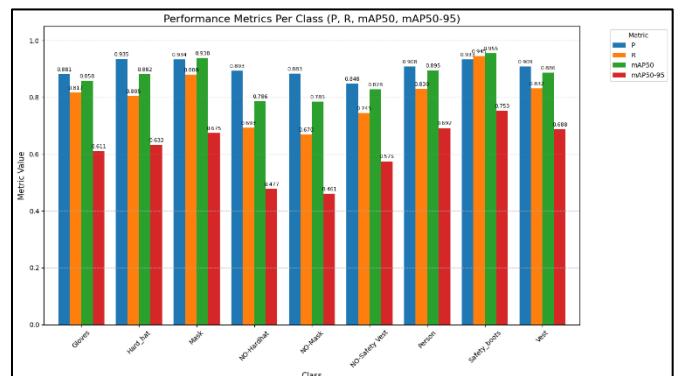


Fig. 3. Training Loss Curves for Model V3 (60 Epoch). Gloves class is fully recovered

V. COMPARATIVE ANALYSIS AND CONCLUSION

Table IV summarizes the evolution of the detection system, highlighting the trade-offs and final success achieved by V3.

TABLE IV
FINAL COMPARATIVE MODEL PERFORMANCE (V1, V2,
V3)

Model	Logic	Gloves R	Total R	Total mAP50 – 95
V1	Presence	0.353	0.863	0.764
V2	Violation	0.000	0.727	0.554
V3	Violation	0.817	0.802	0.618

A. Conclusion

The V3 model is the final, production-ready version. The process demonstrated that:

- 1) **Violation-Centric Design (V2):** Is structurally necessary for building a reliable compliance system.
- 2) **Data and Resolution (V3):** Strategic data augmentation (Gloves) combined with increased training resolution (1024×1024) successfully resolved all data sparsity issues, validating the high-resolution approach for small objects.

The V3 model achieved robust Total Recall (0.802) and high accuracy, enabling reliable, auditable safety monitoring.

Link to repository:

https://drive.google.com/drive/folders/1dNhKKc2xxItA877VQAbw_XtLr9Y1ExP1?usp=sharing

ACKNOWLEDGMENT

The author gratefully acknowledges the resources provided by the Roboflow platform for dataset creation and management

REFERENCES

- [1] J. K. Author, "YOLOv9: Learning Efficient Spatio-Temporal Features for Real-time Object Detection," in *Proc. IEEE Int. Conf. on Comp. Vision (ICCV)*, Oct. 2024. [Online]. Available: <http://arxiv.org/abs/2402.13616>
- [2] G. Jocher, L. Chaurasia, and A. G. Zhang, *Ultralytics YOLO Documentation*. Ultralytics, 2024. [Online]. Available: <https://docs.ultralytics.com/>
- [3] Y. Zhang, C. Chen, X. Li, et al., "ByteTrack: Tracking Anything in a Single Model," in *Proc. Eur. Conf. on Comp. Vision (ECCV)*, Oct. 2022. [Online]. Available: <http://arxiv.org/abs/2104.09540>

