EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Heterogeneous Multi-Model Ensemble for PPE Detection in Construction Environments

Thi-Nguyen Nguyen [1,**], Quang-Anh Nguyen-Duc[2,4,**], Thai Kim Dinh[2,4,*] Duy Tung Doan[3,4], Duc Pham Minh[4], Duc Cuong Van[3], Minh-Anh Nguyen[2,4]

[1]Faculty of Electrical and Electronics Engineering, Vietnam-Hungary Industrial University, Vietnam
[2]International School, Vietnam National University, Hanoi, Vietnam
[3]School of Information and Communications Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
[4]Computer Vision and Robotics Lab, International School, Vietnam National University, Hanoi, Vietnam

Abstract

The construction industry remains one of the most hazardous work environments, with a fatality rate of 25.6 per 100,000 workers, significantly exceeding the average across all industries. PPE compliance is crucial for worker safety, yet monitoring adherence remains challenging in dynamic construction environments. This paper presents an automated PPE detection system utilizing an ensemble deep learning models to enhance workplace safety monitoring. Our approach combines three advanced architectures with Yolov11, RT-DETRv2, and HyperYolo. The individual model predictions are integrated using WBF to improve detection robustness. We evaluate our system on a comprehensive dataset of 4,135 professionally annotated images encompassing critical PPE categories including hard hats, safety vests, protective gloves, and safety boots, along with their corresponding absence classes. The proposed ensemble achieves superior performance with a precision of 0.765, recall of 0.735, mAP@50 of 0.760, and mAP@50:95 of 0.440, outperforming individual models across all evaluation metrics. The results demonstrate the effectiveness of multi-model fusion for automated PPE detection. This research contributes to the advancement of intelligent safety systems that can significantly reduce workplace injuries and fatalities through automated PPE compliance verification.

Received on 16 August 2025; accepted on 09 November 2025; published on 01 December 2025

Keywords: PPE, Detection, Ensemble, Multi-Model, Construction Environments

doi:10.4108/eetiot.9971

1. Introduction

The construction industry remains one of the most hazardous sectors, with significant rates of work-related injuries and fatalities. In 2021, the US Bureau of Labor Statistics reported 5,190 fatal work injuries in the U.S., with nearly 20% occurring in construction [1]. Moreover, a study by Abukhashabah *et al.* has shown that falls account for a significant portion of construction-related fatalities, with falls from outer surfaces comprising 28.36% and falls from buildings 19.39% of such incidents [2]. Additionally, research indicates that the construction sector experiences a high rate of fatal injuries, with a fatality rate of 25.6 per 100,000 full-time workers, which is more than 3.5 times the occupational fatality rate for all industries combined [3]. The utilization of personal protective equipment (PPE) is important in avoiding these hazards; proper use of PPE has been associated with a significant reduction in occupational injuries [4]. Furthermore, studies have demonstrated that the use of PPE can prevent a substantial proportion of workplace injuries, with research indicating that more than 90% of incidents can be mitigated through its proper implementation. However, despite the proven effectiveness of PPE, compliance rates among construction workers remain suboptimal. Factors contributing to this include inadequate training, discomfort, and a lack of enforcement, which collectively diminish the protective benefits of PPE [5], [6].

⋆Corresponding author. Email: thaikd@vnu.edu.vn
⋆⋆Emails: thinnguyen@viu.edu.vn; anhnd@vnuis.edu.vn;

Addressing these challenges necessitates a multifaceted approach, incorporating comprehensive training programs, stringent enforcement of safety protocols, and the integration of advanced technologies. In particular, the application of deep learning algorithms in monitoring PPE compliance offers a promising avenue to enhance workplace safety. Many studies have explored different methods and models to improve the accuracy and efficiency of PPE detection.

Regarding YOLO-based approaches [7], Ludwika and Rifai utilized Yolov5 to detect multiple PPE classes, including helmets of various colors, vests, and safety glasses, achieving an F1-score of 0.744 and a mean average precision (mAP) of 0.757 [8]. Similarly, Wang *et al.* introduced a model based on Yolov5x, which demonstrated superior performance with an mAP of 86.55% across six classes, encompassing helmets of four distinct colors, persons, and vests [9]. Beyond standard implementations, recent research has focused on architectural enhancements. Our previous study [10] introduced a novel approach for detecting PPE in real construction sites using a YOLO backbone integrated with Transformer modules. The proposed architecture was built from the YOLOv5s model by incorporating the SC3T module into the final layer for more accurate detection. Additionally, this model utilized the SIoU loss function instead of the CIoU loss function. More recently, Shi *et al.* introduced GBSG-Yolov8n [11], an enhanced model approach that achieved a 3% improvement over the baseline Yolov8 model. These studies collectively demonstrate a clear trend toward upgrading backbone architectures and refining feature extraction modules to boost PPE detection performance.

In addition to Yolo-based models, other deep learning frameworks have been employed. For example, Ahmed *et al.* developed aFast Region-based Convolutional Neural Network (Fast R-CNN) trained on a dataset of 1,699 annotated images, achieving an mAP of 96% in detecting PPE items such as red, blue, white, and yellow helmets, as well as vests and safety glasses [12]. Furthermore, advancements have been made to enhance the speed and efficiency of PPE detection models. Ke *et al.* developed a lightweight object detection method, reducing model size and resource consumption by 32% and 25%, respectively, with minimal accuracy loss. This approach achieved a detection speed of 105 frames per second (FPS), facilitating deployment on embedded and mobile devices [11].

Despite these advancements, challenges persist in achieving high detection accuracy, particularly in complex environments with varied lighting conditions, occlusions, and diverse PPE types. Most existing works are still constrained by dataset limitations or rely heavily on single-architecture improvements, often lacking the robustness required for dynamic construction sites.

In this research, we continue to focus on enhancing model robustness and generalization capabilities to address these limitations. Our proposed approach provides real-time feedback by automating the identification of PPE usage. It helps ensure compliance with safety regulations, thereby reducing accidents and fatalities in the construction industry. The key contributions of this research are summarized as follows:

(1) We introduce a dataset of 4,135 images specifically designed for construction environment PPE detection, encompassing critical safety equipment categories including *hard hats*, *safety vests*, *protective gloves*, *and safety boots* along with their *corresponding absence classes* to enable comprehensive compliance monitoring in real-world construction scenarios.

(2) We propose a novel heterogeneous ensemble approach that combines diverse architectures (CNN-based Yolov11, Transformer-based RT-DETRv2, and Hypergraph-based HyperYolo) using Weighted Box Fusion to achieve superior PPE detection performance. Our method achieves a precision of 0.765, recall of 0.735, and mAP@50 of 0.760, outperforming individual models across all evaluation metrics.

2. Methodology

2.1. Yolov11

In this study, we adopt Yolov11 [13], a recent advancement in the YOLO (You Only Look Once) object detector family. Yolov11 enhances the one-stage detection paradigm by integrating architectural improvements and training strategies that collectively boost accuracy and inference speed. It introduces a refined backbone with efficient feature extraction capabilities, a redesigned neck structure for more effective multi-scale feature fusion, and a robust detection head optimized for both classification and localization tasks [14]. These enhancements allow Yolov11 to maintain a favorable balance between computational efficiency and detection performance, making it suitable for deployment in resource-constrained and real-time scenarios.

2.2. RT-DETRv2 Model

RT-DETRv2 (Real-Time Detection Transformer v2) [15] is a state-of-the-art end-to-end object detection model designed to achieve high accuracy while maintaining real-time inference speed. Building on the success of RT-DETR [16], this enhanced version incorporates a more efficient transformer-based architecture and an optimized feature extraction pipeline to improve detection performance across diverse scenarios. Unlike conventional anchor-based approaches, RT-DETRv2 leverages a query-based detection framework, where a set of learnable object queries dynamically interact with image features through cross-attention mechanisms, effectively reducing redundant computations. The introduction of dynamic query selection further refines this process by adaptively prioritizing the most relevant queries, enabling precise object localization and classification even in complex and cluttered environments. Additionally, integrating adaptive spatial attention enhances feature aggregation by selectively focusing on the most informative regions, improving detection robustness for small and occluded objects. With an optimized backbone network and a streamlined inference pipeline, RT-DETRv2 significantly reduces latency while maintaining high detection accuracy. It is particularly well-suited for real-time applications such as autonomous driving, surveillance, and industrial automation.

2.3. HyperYolo

To further enhance system robustness, we incorporate Hyper-Yolo, proposed by Feng *et al.* [17]. This is a novel architecture that utilizes hypergraph computation to model complex, high-order correlations among visual features. Unlike traditional grid-based fusion techniques, Hyper-Yolo introduces the Hypergraph Computation Empowered Semantic Collecting and Scattering (HGC-SCS) framework, which transposes feature maps into a semantic space and constructs a hypergraph $G = (V, E)$ over feature vertices V using ϵ-ball based distance thresholds to define hyperedges E. High-order message propagation is then performed via Hypergraph Convolution; this operation is defined by Eq. (1):

$$\text{HyperConv}(X, H) = X + D_v^{-1} H D_e^{-1} H^\top X \Theta, \quad (1)$$

where X is the input feature matrix, H is the incidence matrix, D_v and D_e are the degree matrices of vertices and hyperedges, respectively, and Θ is a trainable weight matrix.

The fused representation X' captures richer semantics and structural context, significantly improving detection robustness. Additionally, Hyper-Yolo introduces the Mixed Aggregation Network (MANet) as a backbone component, combining 1×1 convolutions, depthwise separable convolutions, and C2f-style aggregation to form a diversified feature set as follows:

$$X_{\text{out}} = \text{Conv}_{1\times1}(X_1 \| X_2 \| \cdots \| X_{4+n}), \quad (2)$$

where $\|$ denotes the concatenation operation.

These improvements enable Hyper-Yolo to outperform Yolov8 and Yolov9 across multiple scales with notable gains in AP metrics, especially in lightweight configurations, by explicitly capturing both cross-level and cross-position interactions in the semantic feature space.

2.4. Weighted Box Fusion

Weighted Box Fusion (WBF) is an advanced ensemble technique designed to combine predictions from multiple object detection models to achieve robust performance compared to individual models [**?**]. Unlike traditional ensemble methods such as Non-Maximum Suppression (NMS) or Soft-NMS, WBF leverages the confidence scores of all proposed bounding boxes to generate more accurate and robust predictions.

The fundamental principle of WBF lies in its ability to fuse overlapping bounding boxes from different models by considering their confidence scores as weights. Rather than simply selecting the box with the highest confidence or applying hard suppression, WBF creates averaged boxes that incorporate information from all contributing predictions. This approach is particularly beneficial when different models exhibit complementary strengths, as it can capture the collective intelligence of multiple detection algorithms. The final confidence score for the fused box is computed by Eq. (3):

$$c_{fused} = \frac{\sum_{i=1}^{n} c_i}{n} \cdot \frac{n}{N} \quad (3)$$

where N is the total number of models contributing to the ensemble, and n is the number of boxes that were fused together. The implementation of WBF in this study combines predictions from multiple high-performing detection models, creating a more robust and reliable PPE detection system capable of handling the challenging and diverse conditions typically encountered in construction environments.

2.5. Proposed Architectures

Fig. 1 illustrates the proposed architecture framework. The Training and Validation splits were fed into Yolov11, HyperYolo, and RT-DETRv2 to train for 100 epochs. These models then produced predictions on the test set, which were subsequently ensembled utilizing WBF to achieve robust final results.

3. Experimental Setup

3.1. Dataset

In this research, we introduce a PPE detection dataset tailored for construction environments, comprising a total of 4,135 images collected through web scraping techniques and professionally annotated using the Roboflow platform. The dataset focuses on critical

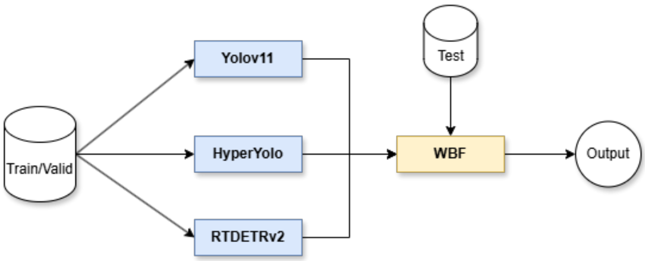

Figure 1. Proposed Architecture

personal protective equipment categories commonly required in construction and industrial settings, including hard hats, safety vests, protective gloves, and safety boots, along with their corresponding absence classes to enable comprehensive compliance monitoring.

The dataset was divided into three subsets following a standard split of 70% for training (2,898 images), 20% for validation (821 images), and 10% for testing (416 images). The distribution of these subsets is outlined in Table 1.

Table 1. Dataset Split Distribution

Split	Images	Percentage
Train	2,898	70.0%
Validation	821	20.0%
Test	416	10.0%
Total	**4,135**	**100.0%**

The dataset exhibits a realistic distribution of PPE compliance scenarios encountered in actual construction sites, with both compliant and non-compliant cases represented across all equipment categories. This balanced representation enables the trained models to effectively learn patterns associated with both proper PPE usage and safety violations. The detailed class distribution is outlined in Table 2.

Table 2. Dataset Class Distribution

Class	Count	Percentage
Person	8,479	21.8%
No Vest	5,643	14.5%
Boots	5,170	13.3%
No Gloves	5,134	13.2%
Hardhat	4,558	11.7%
No Hardhat	3,955	10.2%
Gloves	2,921	7.5%
Vest	2,497	6.4%
No Boots	1,305	3.4%
Total	**38,962**	**100.0%**

3.2. Environment

The experimental setup was conducted on a workstation equipped with an Intel Core i5-13400F processor, 32GB of RAM, and an NVIDIA GeForce RTX 4070Ti SUPER graphics card. The software framework was implemented in Python 3.12.4 utilizing the PyTorch 2.4.0 deep learning library and leveraging CUDA 12.4 for GPU acceleration.

Training was conducted over 100 epochs using a batch size of 16, with input images standardized to 640×640 pixel resolution. The data pipeline employed 8 workers for efficient loading, while mosaic augmentation was disabled after the initial 10 epochs to stabilize training. Model optimization began with an initial learning rate of 0.01, implementing a linear decay strategy, combined with a momentum of 0.937 and weight decay of 0.0005.

3.3. Evaluation Metrics

To comprehensively evaluate the performance of the proposed systems, we employed four standard evaluation metrics widely used in object detection tasks: Precision (P), Recall (R), Mean Average Precision at IoU threshold 0.5 (mAP@50), and Mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP@50:95).

Precision (P) measures the accuracy of positive predictions, calculated as the ratio of true positive detections to the total number of positive predictions made by the model. Recall (R) evaluates the model's ability to identify all relevant objects within the image, computed as the ratio of true positive detections to the total number of actual labels. These metrics are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

where TP represents true positives, FP represents false positives, and FN represents false negatives.

Furthermore, regarding mAP@50, this metric represents the average precision across all PPE classes when detections are considered correct if they have an Intersection over Union (IoU) overlap of at least 50% with the ground truth bounding boxes. The mAP@50:95 is the extended metric calculated by averaging the mAP values across multiple IoU thresholds, ranging from 0.5 to 0.95 in increments of 0.05. These metrics are calculated using Eq. (6) and Eq. (7):

$$mAP@50 = \frac{1}{N}\sum_{i=1}^{N} AP_i@50 \tag{6}$$

$$mAP@50:95 = \frac{1}{10}\sum_{t=0.5}^{0.95} mAP@t \tag{7}$$

where N is the number of PPE classes, $AP_i@50$ is the Average Precision for class i at IoU threshold 0.5, and t represents the IoU threshold values (0.5, 0.55, ..., 0.95). The Average Precision (AP) itself is defined as the area under the Precision-Recall curve:

$$AP = \int_0^1 P(R)\, dR \tag{8}$$

The mAP@50:95 is a stricter metric that emphasizes the importance of precise object localization, as it requires increasingly accurate bounding box predictions to be considered successful detections.

4. Experiments Results

The quantitative performance of the object detection models is summarized in Table 3. The evaluation metrics include Precision (P), Recall (R), mAP@50, and mAP@50:95.

Table 3. Performance Comparison of Object Detection Models

Model	P	R	mAP@50	mAP@50:95
Yolov11x [13]	0.745	0.731	0.759	0.436
RT-DETRv2 [15]	–	–	0.750	0.433
HyperYolo [17]	0.748	0.649	0.699	0.395
WBF Ensemble	**0.765**	**0.735**	**0.760**	**0.440**
Huang and Cheng [19]	0.589	0.730	0.691	0.483
Nguyen *et al.* [18]	0.740	0.650	0.680	–

As presented in Table 3, the proposed WBF ensemble, which integrates predictions from Yolov11x, RT-DETRv2, and HyperYolo, achieves the highest performance among the tested configurations, with a Precision of 0.765, Recall of 0.735, mAP@50 of 0.760, and mAP@50:95 of 0.440.

Analyzing the individual models reveals distinct characteristics: Yolov11x serves as a strong baseline with balanced metrics. HyperYolo demonstrates competitive Precision (0.748) but suffers from a lower Recall (0.649), indicating it misses some objects in complex scenes. However, by applying the Weighted Box Fusion (WBF) technique, we effectively mitigate this limitation. The ensemble leverages the high recall of Yolov11x and the structural robustness of HyperYolo and RT-DETRv2, resulting in a fused model that outperforms any single constituent model across key metrics.

Compared to previous studies, our WBF ensemble consistently outperforms existing approaches such as Nguyen *et al.* [18] in terms of mAP@50 (0.760 vs 0.68) and shows significant improvement in Precision compared to Huang and Cheng [19] (0.765 vs 0.589). These results demonstrate the effectiveness of the WBF ensemble in synergizing the strengths of individual models to enhance detection robustness in construction environments.

Furthermore, Fig. 2 illustrates qualitative inference results of the proposed architectures. As observed, the system effectively detects PPE items such as hard hats, vests, and gloves even under challenging conditions, including cluttered backgrounds, varying lighting, and partial occlusions. The bounding boxes are tightly aligned with the objects, confirming the localization accuracy of the ensemble method.

5. Conclusion

This study presents a comprehensive approach to automated PPE detection in construction environments using an ensemble of deep learning models. Our research addresses the critical need for improving workplace safety in the construction industry, where inadequate PPE compliance contributes significantly to the high rates of occupational injuries and fatalities. The proposed methodology combines three advanced detection architectures: Yolov11, RT-DETRv2, and HyperYolo. These were integrated through WBF to leverage the unique strengths of each model.

Our experimental results demonstrate that the ensemble approach achieved higher performance compared to individual models, with the WBF ensemble attaining a precision of 0.765, recall of 0.735, mAP@50 of 0.760, and mAP@50:95 of 0.440. These results represent meaningful improvements over the individual model performances, with Yolov11x achieving competitive results as a standalone solution.

In conclusion, our ensemble-based approach showed significant potential for automating PPE detection in construction environments. The integration of multiple detection paradigms through WBF provides a robust foundation for practical deployment, contributing to the broader goal of reducing workplace injuries and fatalities through intelligent safety monitoring systems. As the construction industry continues to embrace digital transformation, automated PPE detection systems like the one presented in this study will play an increasingly vital role in creating safer work environments and protecting worker welfare.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgment

This research was funded by the International School, Vietnam National University, Hanoi. Gratefully acknowledge the ATRC for their valuable support in the design of the figures.

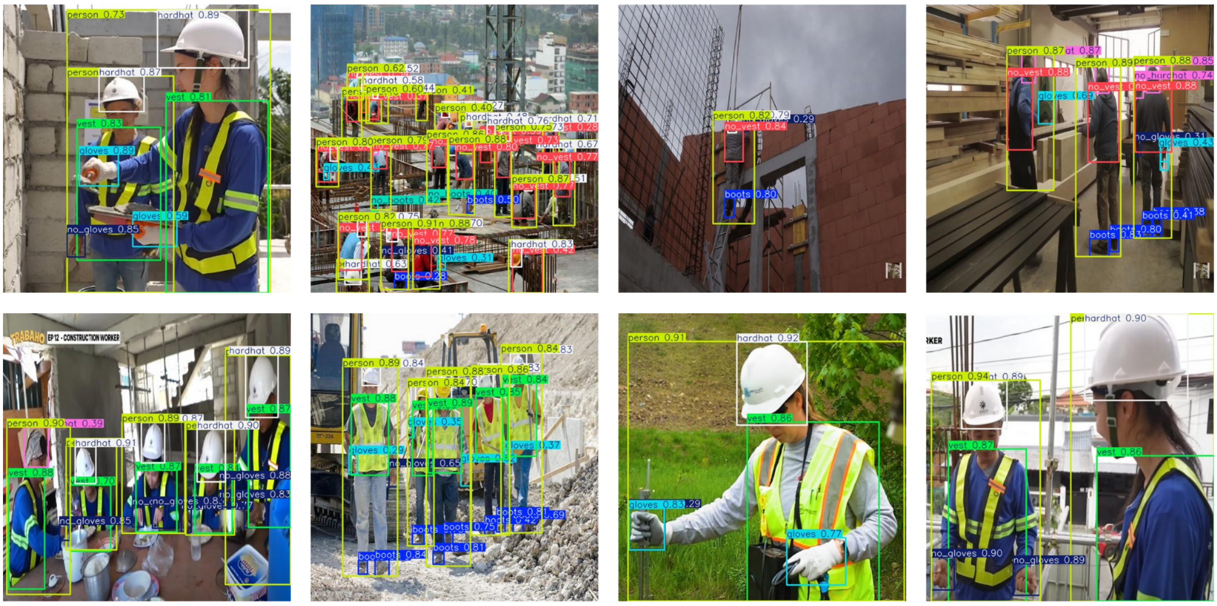

Figure 2. Inference Results From Proposed Architecture

References

[1] Bureau of Labor Statistics, U.S. Department of Labor. (2023). Construction deaths due to falls, slips, and trips increased 5.9 percent in 2021. Available at: https://www.bls.gov/opub/ted/2023/construction-deaths-due-to-falls-slips-and-trips-increased-5-9-percent-in-2021.htm (accessed March 24, 2025).

[2] Abukhashabah, E., Summan, A., & Balkhyour, M. (2020). Occupational accidents and injuries in construction industry in Jeddah city. *Saudi Journal of Biological Sciences*, 27(8), 1993–1998.

[3] Kisner, S. M., & Fosbroke, D. E. (1994). Injury hazards in the construction industry. *Journal of Occupational and Environmental Medicine*, 36(2), 137–143.

[4] Kursunoglu, N., Onder, S., & Onder, M. (2022). The evaluation of personal protective equipment usage habit of mining employees using structural equation modeling. *Safety and Health at Work*, 13(2), 180–186.

[5] Khoshakhlagh, A. H., Malakoutikhah, M., Park, J., Kodnoueieh, M. D., Boroujeni, Z. R., Bahrami, M., & Ramezani, F. (2024). Assessing personal protective equipment usage and its correlation with knowledge, attitudes, performance, and safety culture among workers in SMEs. *BMC Public Health*, 24(1), 1987.

[6] Alemu, A. A., Yitayew, M., Azazeh, A., & Kebede, S. (2020). Utilization of personal protective equipment and associated factors among building construction workers in Addis Ababa, Ethiopia. *BMC Public Health*, 20, 1–7.

[7] Lavanya, G., & Pande, S. D. (2024). Enhancing real-time object detection with YOLO algorithm. *EAI Endorsed Transactions on Internet of Things*, 10.

[8] Ludwika, A. S., & Rifai, A. P. (2024). Deep learning for detection of proper utilization and adequacy of personal protective equipment in manufacturing teaching laboratories. *Safety*, 10(1), 26.

[9] Wang, Z., Wu, Y., Yang, L., Thirunavukarasu, A., Evison, C., & Zhao, Y. (2021). Fast Personal Protective Equipment Detection for Real Construction Sites Using Deep Learning Approaches. *Sensors*, 21(10), 3478.

[10] Do, M. T., Kim, T. D., Ha, M. H., Chen, O. T. C., Nguyen, D. C., & Tran, A. L. Q. (2023, December). An Effective Method for Detecting Personal Protective Equipment at Real Construction Sites Using the Improved YOLOv5s with SIoU Loss Function. In *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)* (pp. 430–434). IEEE.

[11] Ke, X., Chen, W., & Guo, W. (2022). 100+ FPS detector of personal protective equipment for worker safety: A deep learning approach for green edge computing. *Peer-to-Peer Networking and Applications*, 15(2), 950–972.

[12] Ahmed, M. I. B., Saraireh, L., Rahman, A., Al-Qarawi, S., Mhran, A., Al-Jalaoud, J., Al-Mudaifer, D., Al-Haidar, F., AlKhulaifi, D., Youldash, M., & Gollapalli, M. (2023). Personal Protective Equipment Detection: A Deep-Learning-Based Sustainable Approach. *Sustainability*, 15(18), 13990.

[13] Jocher, G., Qiu, J., & Chaurasia, A. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. `https://github.com/ultralytics/ultralytics`

[14] Khanam, R., & Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725.

[15] Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., & Liu, Y. (2024). RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer. arXiv preprint arXiv:2407.17140.

[16] 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).

[17] Feng, Y., Huang, J., Du, S., Ying, S., Yong, J. H., Li, Y., ... & Gao, Y. (2024). Hyper-YOLO: When visual object detection meets hypergraph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[18] Nguyen, N. T., Tran, Q., Dao, C. H., Nguyen, D. A., & Tran, D. H. (2024). Automatic detection of personal protective equipment in construction sites using metaheuristic optimized YOLOv5. *Arabian Journal for Science and Engineering*, 49(10), 13519–13537.

[19] Huang, M. L., & Cheng, Y. (2025). Dataset of personal protective equipment. *Data in Brief*, 111988.