A Lightweight Safety Helmet Compliance Detection via Multimodal Fusion (Student Abstract)

Jeong Hwan Ryu*[1], Azimjon Akhtamov*[1], Md Azher Uddin†[2], Aziz Nasridinov†[1]

[1]Department of Computer Science, Chungbuk National University, Cheongju, 28644, South Korea
[2]School of Mathematical and Computer Science, Heriot-Watt University, Dubai 501745, United Arab Emirates
{ryujh030820, azimjan21, aziz}@chungbuk.ac.kr, m.uddin@hw.ac.uk

Abstract

Ensuring proper use of personal protective equipment (PPE), especially helmets, is essential for workplace safety. Conventional object detectors often fail to distinguish whether a helmet is worn correctly, and existing approaches relying on single-model pipelines are prone to localization errors and false alarms. Moreover, most prior studies do not guarantee real-time performance. To resolve these challenges, we propose a lightweight multimodal approach that integrates a YOLO11-based object detector with a pose estimation model, achieving higher F1 scores and lower false alarm rates while maintaining real-time performance.

Introduction

Recently, artificial intelligence (AI)–based monitoring has gained attention as a way to verify PPE compliance. In particular, several studies have attempted to monitor PPE compliance using object detection models such as YOLO (You Only Look Once). For instance, Zhou et al. (2021) applied YOLOv5 for real-time helmet detection. However, conventional object detection models often struggle to determine whether a helmet is being properly worn, such as distinguishing between a helmet held in hand and one actually worn on the head.

Keypoint-based methods have been explored to address object-detection limitations by providing precise information about human body positions. For example, Chen and Demachi (2021) matched PPE items to body parts using geometric rules and full-body pose information. However, their method uses an approach in which pose estimation and object detection are performed by separate models, such as OpenPose and YOLOv3, which may increase computational cost. More recently, Akhtamov et al. (2025) presented a fusion-based approach that uses both object detection results and pose information to improve PPE compliance detection. However, this method still relies on separate models and was not designed for real-time performance.

To overcome these limitations, we propose a lightweight multimodal approach that jointly performs object detection and pose estimation for accurate helmet-wearing compliance detection. First, our model adopts a shared-backbone structure and simplified architecture, enabling both tasks to reuse common features more efficiently. This design reduces redundant computation and ensures consistent real-time performance. Second, we incorporate a pose-guided distance filtering module that evaluates the geometric relationship between helmet bounding boxes and head keypoints. This module effectively removes mismatches and reduces false alarms.

Proposed Method

Our approach consists of three main components: a shared backbone, task-specific detection and pose branches, and a pose-guided fusion module, as illustrated in Figure 1.

Shared Backbone. To achieve real-time performance, the proposed architecture integrates object detection and pose estimation into a single lightweight design. Both tasks share the same YOLO11 backbone, enabling them to reuse common visual features across tasks. To further reduce inference cost, we modify the backbone by replacing all C3K2 modules with RepC3 blocks. These blocks learn richer multi-branch features during training and later collapse into a single 3×3 convolution for faster inference. The final backbone layer remains unchanged to maintain a balance between accuracy and model size.

Object Detection and Pose Branches. The detection branch focuses on locating helmets, while the pose branch predicts head keypoints. Both components are simplified to minimize computation. In the object detection branch, only the P3 head is retained since helmets are small objects, while the P4 and P5 heads intended for medium and large targets are removed. In the pose estimation branch, the P5 head is removed to eliminate redundant large-scale processing, which reduces the number of parameters and inference time. The model predicts only five head-related keypoints (nose, eyes, and ears), which are sufficient for helmet localization and compliance analysis while keeping the architecture lightweight. The neck structure is also simplified by replacing C3K2 modules with RepC3 blocks for consistent feature extraction and reduced computation across both branches.

*These authors contributed equally.
†The Corresponding authors.

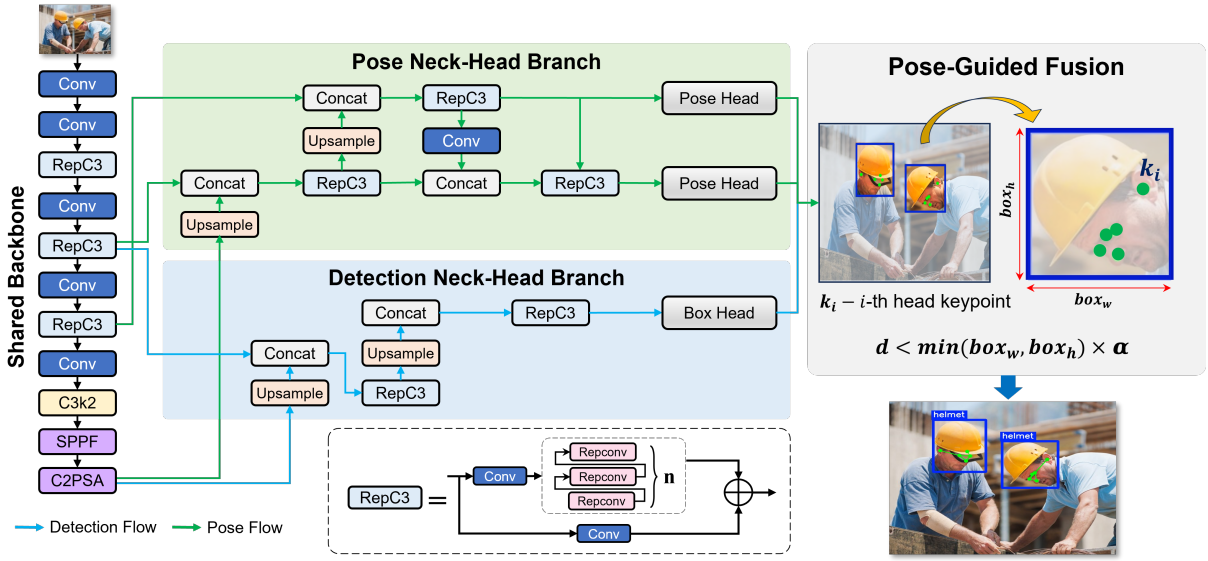

Figure 1: The proposed framework includes a shared backbone, detection and pose modalities, and a pose-guided fusion module.

Pose-Guided Fusion. To accurately verify helmet-wearing compliance, only helmet detections paired with a corresponding head keypoint group are considered valid. For each detected helmet, the distance d to the nearest keypoint k_i is computed as

$$d < \min(box_w, box_h) \times \alpha,$$

where α is a scaling constant experimentally set to 0.4. If a keypoint is within the helmet bounding box, d is set to 0. This geometric constraint maintains spatial consistency between the helmet and head region, reducing false alarms.

Experimental Results

Dataset. For training the object detection model, we used the CPPE dataset (Xiong et al. 2021), excluding manually annotated samples in which helmets were improperly worn. For training the pose estimation model, we employed the COCO 2017 keypoint dataset, retaining only the five head-related keypoints and discarding the remaining ones.

Training Details. All models were trained on a single NVIDIA TITAN RTX GPU. Training applied a batch size of 16 using 640×640 images, a 0.01 learning rate, and 0.0005 weight decay.

Result. Table 1 shows the comparison results between the baseline models and our proposed method. Our model outperforms both YOLO11n and YOLO11s in accuracy, achieves fewer false positives, and demonstrates higher FPS performance by utilizing a shared backbone and head keypoint guidance for precise and efficient detection.

Conclusion and Future Work

In this study, we proposed a helmet detection model that improves accuracy while maintaining real-time, lightweight performance. Compared to YOLO11n and YOLO11s, our approach achieves higher FPS and a lower false alarm rate. Future work includes enhancing the pose module by fine-tuning it on workplace-specific data to boost model reliability.

Method	Precision	Recall	F1	FPS	False Positive (FP)
YOLO11n	0.896	0.959	0.927	81.7	91
YOLO11s	0.926	**0.964**	0.944	65.7	63
Our method	**0.955**	0.938	**0.946**	**110.7**	**36**

Table 1: Result of experiments (hardhat only).

Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-25443739, DeepHunter: Evolutionary Deepfake Detection Technology for Identifying Adversarial Synthetic Images).

References

Akhtamov; et al. 2025. A Multimodal Fusion Model for Enhanced Industrial Glove-Wearing Compliance Detection. *Proceedings of the AAAI Symposium Series*, 6(1): 75–77.

Chen, S.; and Demachi, K. 2021. Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. *Automation in Construction*, 125: 103619.

Xiong; et al. 2021. Pose guided anchoring for detecting proper use of personal protective equipment. *Automation in Construction*, 130: 103828.

Zhou; et al. 2021. Safety helmet detection based on YOLOv5. In *2021 IEEE International conference on power electronics, computer applications (ICPECA)*, 6–11. IEEE.