A Multimodal Fusion Model for Enhanced Industrial Glove-Wearing Compliance Detection

Azimjon Akhtamov[1], Aziz Nasridinov[1, *], Sang Hyun Choi[2, *]

[1]Department of Computer Science, Chungbuk National University, Cheongju, 28644, South Korea
[2]Department of Management Information Systems, Chungbuk National University, Cheongju, 28644, South Korea
{azimjan21, aziz, chois}@chungbuk.ac.kr

Abstract

Glove detection in manufacturing environments is challenging due to glove-background blending and limited dataset diversity. To address this, we propose a multimodal detection framework that enhances segmentation models through wrist keypoint-guided feature fusion, effectively reducing false negatives. We also introduce a unified dataset spanning five manufacturing domains to improve generalizability. Experimental results show our method achieves mAP 0.821, outperforming the baseline YOLOv11-Seg (mAP 0.792). This highlights the effectiveness of feature fusion between segmentation and keypoints for accurate and reliable glove compliance monitoring in industrial settings.

Introduction

Personal protective equipment (PPE) plays a vital role in ensuring the safety of workers, particularly in industrial environments. The Occupational Safety and Health Administration (OSHA) reports that 71% of hand injuries could be prevented with proper hand protection, specifically safety gloves. However, a significant issue persists, with 70% of workers not wearing gloves at all, and 30% of those who do wear the incorrect type for their specific tasks. This highlights the importance of accurate and reliable glove detection systems to ensure workers stick to safety protocols.

Existing studies have mostly used object detection for glove detection in industrial settings. For example, Gugssa et al. (2021) proposed YOLO for the direct detection of gloves and YOLO with a CNN for usage classification. Yu et al. (2023) introduced a YOLO-based system for electric sites using a custom attention module and transfer learning. Li et al. (2023) and Tao et al. (2024) enhanced YOLOv8 for machinery workshops and glove compliance in power grids. However, bounding box-based methods often include irrelevant pixels and lack boundary precision, leading to false detections, especially as gloves are small, vary in appearance, and blend into backgrounds, which raises false alarms. Moreover, existing works are domain-specific and lack a unified dataset for general glove detection.

This paper proposes a novel multimodal model for glove detection that tackles boundary accuracy and false alarm reduction. Unlike previous methods, it uses instance segmentation for pixel-level glove detection, overcoming bounding-box limitations. A wrist keypoint-guided matching module activates when segmentation fails due to background blending or glove variation, reducing false positives. The architecture fuses segmentation and keypoint features for robust detection in industrial settings. A unified dataset spanning industries, glove types, and environments improves generalizability. Experimental results show higher accuracy and fewer false negatives than the baseline.

Proposed Method

Unified Dataset. A unified dataset of 2,391 images was created across five critical industries where gloves are mandatory: construction (521), chemical handling (491), electrical operation (451), metalworking and welding (451), and heavy machinery (477). It includes glove-related images from online repositories and the SH17 (Ahmad & Rahimi, 2024) dataset, which focuses on PPE detection and provides glove images annotated for object detection. These images were re-annotated for segmentation tasks using the Roboflow annotation tool.

Modalities. Figure 1 illustrates the proposed multimodal fusion model for glove detection. The purpose of each modality is to improve glove detection and minimize false alarms. First, we used the YOLOv11-Seg model, which performs pixel-level segmentation to generate detailed gloves and

*The corresponding authors.

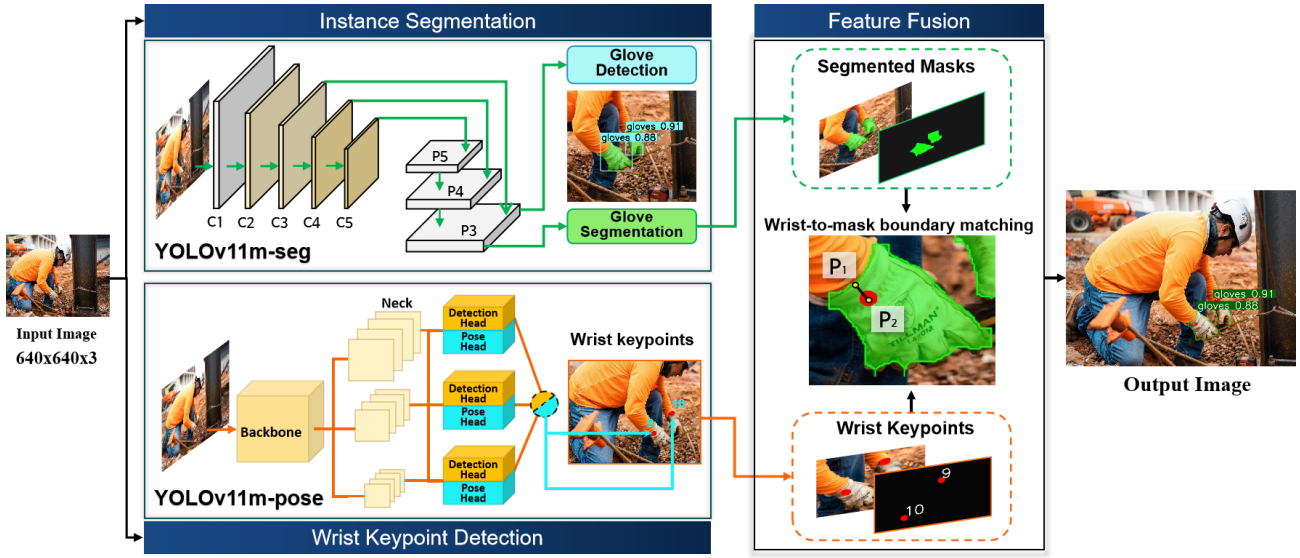

Figure 1: The proposed method includes two modalities: glove segmentation (YOLOv11m-seg) and wrist keypoint detection (YOLOv11m-Pose), fused via masked keypoints.

no gloves masks along with bounding boxes. Only the segmented masks are extracted, providing a detailed representation of glove regions without unnecessary background information. This enables high-precision localization, particularly in cases where gloves are partially occluded, vary in shape or material, or blend into the background. Then, YOLOv11m-Pose is used to detect workers by identifying the person class with its corresponding bounding boxes and full-body keypoints. From these outputs, only the wrist keypoints (denoted as 9 and 10 in the skeleton structure) are extracted, enabling our method to precisely track hand positions while ignoring irrelevant joints.

Feature Fusion. We aggregate segmented glove masks and wrist keypoints using a proximity-based wrist-to-mask matching. Each detected wrist keypoint (P1) is associated with the nearest side of the glove polygon (P2) by evaluating spatial distance and positional constraints to make the pairings hand-wise consistent.

Experimental Results

Training details. The YOLOv11-Seg model was trained on a TITAN RTX GPU for 100 epochs using 640-pixel images and a batch size of 16. The dataset was split 80/20 (1905/486 images), and training used AdamW with a 0.001 learning rate and cosine decay.

Comparative Results. Table 1 shows the multimodal model outperforming baseline (i.e., YOLOv11-Seg), with Recall up 30.3%, F1-Score 17.6%, and mAP 21.3%, highlighting wrist keypoint matching's effectiveness. These improvements are due to our multimodal approach, which fuses segmented masks with wrist keypoints to enhance glove localization and reduce false alarms from background blending, as illustrated in Figure 2.

Model	Precision	Recall	F1	mAP		mAP
				gloves	no-gloves	
Baseline	0.868	0.667	0.75	0.697	0.887	0.792
Our	**0.906**	**0.86**	**0.88**	**0.84**	**0.88**	**0.821**

Table 1: Result of experiments.

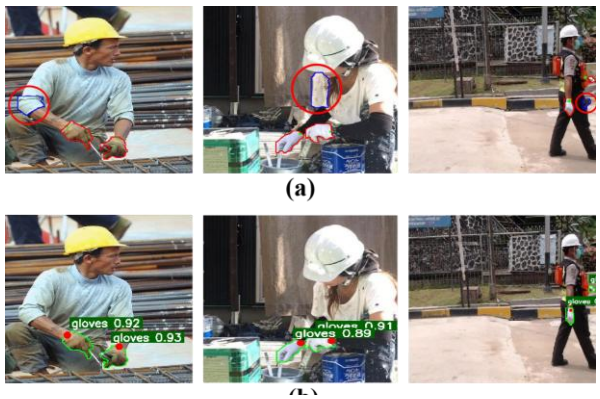

Figure 2: Model performance comparison: (a)Baseline model with false alarms, (b) Our model.

Conclusion and Future Work

This study proposes a multimodal glove detection model using instance segmentation with keypoint-based wrist matching to improve accuracy and reduce false alarms. The model effectively minimizes false positives in complex industrial settings. In the future, we plan to integrate temporal tracking for dynamic safety monitoring and enhance our unified dataset.

References

Ahmad, H. M.; Rahimi, A. 2024. SH17: A dataset for human safety and personal protective equipment detection in the manufacturing industry. arXiv:2407.04590.3.

Gugssa, M.; Ali, G.; Jun, W.; Junfeng, M.; Joshua, B. 2021. PPE-Glove Detection for Construction Safety Enhancement Based on Transfer Learning. In *Proceedings of the ASCE International Conference on Computing in Civil Engineering,* DOI:10.1061/9780784483893.008

Li, S.; Huang, H.; Meng, X.; Wang, M.; Li, Y.; Xie, L.2023. A Glove-Wearing Detection Algorithm Based on Improved YOLOv8. *Sensors*, *23*(24), 9906.

Northwestern University. (2024, March). Volume 8, Issue 3 - March 2024. *Environmental Health & Safety Training Newsletter.* https://www.northwestern.edu/environmental-health-safety/training/newsletters/volume-8-issue-3-march-2024.html

Tao, C.; Wang, C.; Li, T. 2024. Detection research of insulating gloves wearing status based on improved YOLOv8s algorithm. *Journal of Engineering and Applied Science,* 71, 126.

Yu, F.; Zhu J.; Chen, Y.; Liu, S.; Jiang, M. 2023. CAPN: a Combine Attention Partial Network for glove detection. *PeerJ Computer Science*, 9:e1558