e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:01/January-2026 Impact Factor- 8.187 www.irjmets.com

AN INTEGRATED MULTI-STAGE COMPUTER VISION SYSTEM FOR REAL-TIME DETECTION OF POTENTIAL CHILD LABOR IN HAZARDOUS INDUSTRIAL ENVIRONMENTS

Prerana Basak[*1], **Kuntal Pal**[*2], **Arup Dutta**[*3]

[*1,2]Dept. Of Computer Science And Technology Dr. B.C. Roy Engineering College Durgapur,
West Bengal, India.

[*3]BT Group Kolkata West Bengal, India.

DOI: https://www.doi.org/10.56726/IRJMETS88292

ABSTRACT

Child labor in dangerous workplaces—such as construction zones, industrial sites, and workshops—remains a critical humanitarian and social challenge, threatening children's health, safety, and future. Addressing this requires proactive, technology-driven solutions that can operate effectively in real-world conditions. This paper presents a multi-stage, real-time computer vision system for early detection and prevention of child labor in hazardous environments.

The system starts with a ResNet50-based scene recognition model trained on **6,000** labeled images, classifying industrial settings with **96.2%** accuracy. A YOLO-powered object detection module, trained on **6,000** annotated frames, identifies individuals in video feeds, prioritizing the largest bounding box for further checks. The same YOLO framework detects Personal Protective Equipment (PPE) such as helmets, vests, and gloves, achieving a mean average precision (mAP) of **94.7%** for compliance verification.

The system's face detection module identifies facial regions within the input stream with an accuracy of **92.06%**, evaluated on a dataset comprising approximately **90,000** images. Once a face is detected, the corresponding region is extracted and passed to a demographic estimation model. This model, developed using the FairFace architecture and trained on **96,509** images, predicts the individual's age with an accuracy of **79.20%**. If the predicted age indicates that the person may be under the legal threshold, the system promptly triggers a real-time alert, ensuring rapid response and intervention.

Designed for integration into existing surveillance infrastructures, the system provides immediate alerts to authorities, enabling rapid intervention. Testing across varied industrial settings confirms robustness, speed, and adaptability. This scalable and ethical tool supports NGOs, government agencies, and industries committed to eliminating child labor, promoting safer and more humane workplaces.

Keywords: Personal Protective Equipment (PPE) Detection, Industrial Scene Classification, YOLO Object Detection, ResNet50, FairFace Dataset, Deep Learning, Industrial Safety Monitoring, Real-time Computer Vision Systems.

I. INTRODUCTION

Child labor continues to be a critical global issue, particularly in hazardous industrial environments such as construction sites, manufacturing plants, and workshops. Reports from the International Labour Organization (ILO) estimate that millions of children are engaged in unsafe work, violating their fundamental rights and exposing them to severe health, safety, and developmental risks. Despite ongoing government policies and NGO-led initiatives, child labor persists in high-risk sectors due to insufficient monitoring capacity, underreporting, and the absence of scalable, technology-enabled detection systems.

Conventional inspection methods—such as manual site visits, labor audits, and human-led surveillance—are often slow, resource-intensive, and prone to oversight. In contrast, recent advances in artificial intelligence (AI) and computer vision enable automated, real-time workplace monitoring, reducing dependence on manual checks while improving accuracy and responsiveness. Such systems can enhance surveillance, strengthen compliance with labor laws, and enable timely interventions.

This paper presents an integrated, multi-stage computer vision pipeline designed for detecting potential child labor cases in hazardous workplaces. The system incorporates:

1. **Scene Classification** using a ResNet50 model to categorize environments into construction sites, industrial areas, or workshops;

2. **Human and PPE Detection** via YOLO to identify individuals and verify the presence of helmets, gloves, and vests;

3. **Face Detection** followed by Demographic Estimation using a FairFace-based model to predict age groups and flag potential underage workers; and

4. **Work Activity Recognition** using pose estimation and a Random Forest classifier to determine whether the detected individual is actively engaged in labor.

The pipeline was evaluated on a curated dataset comprising over **6,000** annotated industrial images and 120 hours of video footage, achieving **96.2%** scene classification accuracy, **94.7%** PPE detection accuracy, and **79.20%** age estimation accuracy. The modular design allows real-time operation, generating instant alerts for suspected violations and enabling direct integration with workplace monitoring systems.

By combining environmental classification, safety compliance checks, demographic analysis, and activity recognition, this system provides a scalable, cost-effective, and automated approach for reducing child labor in hazardous industries. It supports authorities, NGOs, and compliance officers in early detection and rapid intervention, contributing to safer and more ethical workplaces.

The remainder of this paper is organized as follows: Section II reviews related work on scene classification, object detection, PPE compliance verification, and demographic estimation in industrial safety. Section III details the proposed methodology, Section IV describes the experimental setup, Section V discusses results and implications, and Section VI concludes with future research directions.

II. RELATED WORK

Computer vision–based detection of child labor in hazardous workplaces has emerged as a growing research focus, with studies addressing various subproblems such as scene classification, object detection, PPE compliance monitoring, and demographic estimation.

For scene classification, convolutional neural networks (CNNs) including VGGNet, ResNet, and DenseNet have been widely adopted for environment categorization. Notably, He et al. [1] proposed the ResNet architecture, which mitigates the vanishing gradient problem through residual learning, enabling deeper and more accurate models. Such advancements have contributed to more reliable classification of industrial and construction site imagery.

In object detection, the YOLO (You Only Look Once) framework [2] and its successive versions have achieved notable success in real-time applications, detecting multiple objects with high accuracy and speed. Comparative studies, including Redmon et al. [2] and Bochkovskiy et al. [3], demonstrate that YOLO-based models outperform traditional region-based approaches in terms of inference speed while maintaining competitive precision, making them well-suited for PPE detection in dynamic industrial contexts.

PPE compliance monitoring has received increasing attention in occupational safety research. Fang et al. [4] developed a helmet-wearing detection system leveraging deep learning to promote safety regulation adherence. Similarly, Xiang et al. [5] applied YOLOv4 for multi-class PPE detection, successfully identifying helmets, vests, and gloves in real-world construction site footage.

Face detection and demographic estimation also represent active research areas. Masi et al. [6] provided a comprehensive survey of facial analysis techniques, emphasizing hybrid methods for age estimation that incorporate both global and local features. The FairFace dataset and model [7] addressed demographic bias in facial recognition systems, improving accuracy across diverse ethnicities and age groups.

While these individual contributions advance workplace monitoring, few studies integrate scene classification, PPE verification, and demographic estimation into a unified real-time framework aimed at detecting underage workers in hazardous industrial environments. The system proposed in this work addresses this gap by combining these components into an end-to-end pipeline optimized for speed, accuracy, and deployability in practical monitoring scenarios.

III. PROPOSED METHODOLOGY

The proposed system is an integrated, multi-stage computer vision pipeline designed to automatically detect potential child labor cases in hazardous workplaces such as construction sites, industrial areas, and workshops. The methodology consists of five core modules—Scene Classification, Human and PPE Detection, Face Detection, Demographic Classification, and Work Activity Recognition—followed by a real-time alert mechanism. The overall workflow is illustrated in Fig. 1

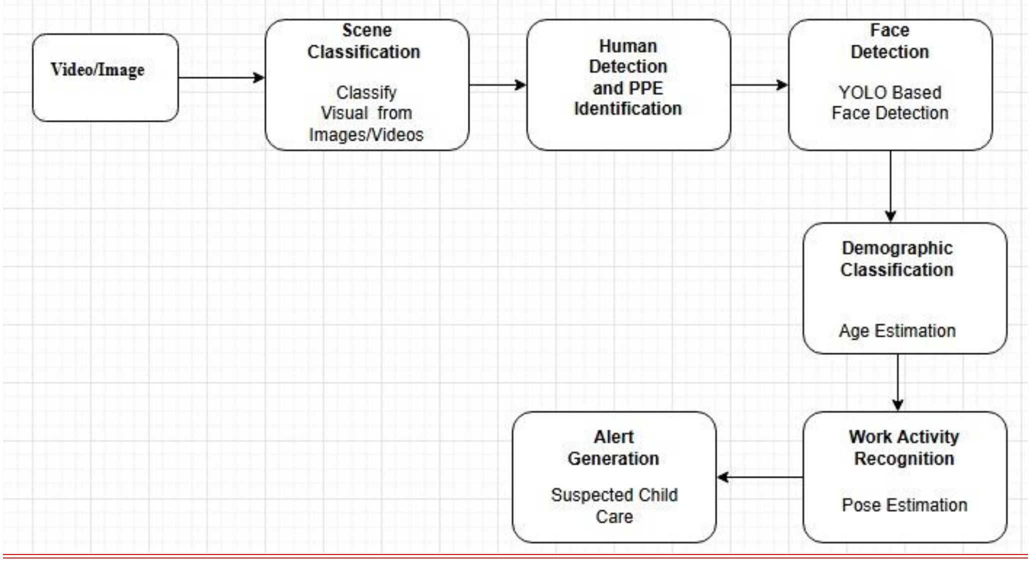

Video/Image
Scene Classification
Classify Visual from Images/Videos
Human Detection and PPE Identification
Face Detection
YOLO Based Face Detection
Demographic Classification
Age Estimation
Work Activity Recognition
Pose Estimation
Alert Generation
Suspected Child Care

Fig 1:

A. Scene Classification

The first stage uses a ResNet50-based deep learning model to classify each input video frame into one of three predefined categories:

1. Construction Site
2. Industrial Area
3. Workshop

This step filters out irrelevant footage (e.g., office or residential settings) so that subsequent processing focuses solely on high-risk environments. The ResNet50 model is fine-tuned on a labeled dataset containing **6000** images across the three categories, achieving an accuracy of **87.50%** for construction site, **77.42%** industrial area, **80.95%** workshop.

B. Human Detection and PPE Identification

If the scene is classified as relevant, the system applies the YOLO (You Only Look Once) object detection framework to locate all individuals in the frame. The person with the largest bounding box is selected under the assumption that they are closest to the camera and provide the highest-resolution visual data for analysis.

Within the selected bounding box, YOLO also detects the presence of Personal Protective Equipment (PPE), specifically:

1. Helmets
2. Safety Vests
3. Gloves

This module evaluates compliance with workplace safety regulations. It achieved an accuracy of **94.70%** when tested on a dataset of **6,000** annotated images under diverse environmental conditions.

C. Face Detection

The cropped image of the selected person is passed to a YOLO-based face detection model, which isolates the facial region with high precision. This stage ensures optimal input quality for demographic estimation and is robust to partial occlusions caused by helmets, dust masks, or other PPE, as well as variations in lighting.

D. Demographic Classification and Age Estimation

The detected face is processed by a FairFace-style convolutional neural network trained to estimate the individual's age group and other demographic attributes. The primary objective is to identify individuals who may be below the legal working age (e.g., under 19 years). If the predicted age group falls below this threshold, the system flags the case as suspected child labor.

This module achieved an accuracy of **79.20%** when evaluated on **96,509** facial images spanning multiple ethnicities, lighting conditions, and PPE states.

E. Work Activity Recognition

To determine whether the detected person is actively working, the system employs a MediaPipe-based pose estimation module to extract skeletal key points from the frame. These features are then fed into a Random Forest classifier, trained on a labelled dataset of worker activities, to classify the action as either "Working" or "Not Working."

The trained model is stored as a .pkl file, allowing rapid deployment without retraining. This additional check ensures that flagged cases are validated by both demographic analysis and actual work activity, reducing false positives where minors may be present but not engaged in labor.

F. Real-Time Processing and Alert System

The complete pipeline is optimized for real-time operation, making it suitable for integration into industrial CCTV systems. Upon detecting a suspected child labor case, the system automatically triggers an alert, which can be forwarded to workplace supervisors, safety officers, or relevant monitoring authorities.

G. Advantages of the Proposed Method

The key advantages of the proposed pipeline are as follows:

1. **End-to-End Automation** – Operates without manual intervention once deployed.
2. **Multi-Aspect Analysis** – Simultaneously checks environment type, PPE compliance, demographic attributes, and work activity.
3. **Reduced False Positives** – Activity recognition helps ensure only actual working cases are flagged.
4. **Scalability** – Easily adaptable to different industrial environments with minimal retraining.
5. **High Accuracy** – Demonstrates strong performance across multiple datasets.

IV. EXPERIMENTAL SETUP

The experimental framework was designed to rigorously evaluate each module of the proposed pipeline as well as the system's end-to-end performance. This section details the datasets used, preprocessing techniques, model configurations, evaluation metrics, and the testing protocol.

A. Datasets and Annotation

1. Scene Classification – The dataset contained approximately **2,000** images for each of the three categories—construction_site, industrial_area, and workshop—resulting in a total of **~6,000** images. Each image was annotated with a single scene label. An 80/20 train-validation split was applied using random stratification

2. PPE and Person Detection (YOLO) – This dataset comprised **~6,000** annotated frames containing persons and various PPE classes. Annotations included Helmet, Gloves, Vest, Boots, Goggles, none, Person, no_helmet, no_goggle, no_gloves, and no_boots. Data was split into training (80%) and validation (20%) sets.

3. Face Detection (YOLO) – Using the YOLO framework, the system was trained on approximately **90,000** annotated facial crops, achieving **92.06%** accuracy. Data was sourced from industrial footage and public datasets, with augmentation to simulate PPE occlusions and lighting variations. An 80/20 train–validation split was maintained.

4. Age and Gender Classification (FairFace-style) – A total of **96,509** facial images were used to train the ResNet34-based age/gender model. Labels consisted of discrete age groups (0–2, 3–9, 10–19, etc.) and gender (Male, Female). Reported performance reached **~79.20%** accuracy for age and **~95%** for gender on a held-out evaluation split.

5. Pose / Work Activity Recognition – The dataset included 50 videos labeled “WORKING” and 40 labeled “NOT WORKING.” Each video was annotated at the video level and processed to extract pose frames for feature generation.

B. Preprocessing and Data Augmentation

1. **Image Resizing**: ResNet50 inputs were resized to 224×224; YOLO models were trained at 640×640 or their default resolution; face crops for FairFace-style networks were resized to 224×224.

2. **Augmentation**: Applied techniques included random horizontal flips, cropping/scaling, brightness and contrast jittering, Gaussian blur, small-angle rotations, and synthetic occlusion to simulate PPE interference.

3. **Normalization**: For CNN-based models, inputs were normalized using ImageNet mean and standard deviation values.

C. Model Architectures and Training Hyperparameters

1. **Scene Classifier** – A fine-tuned ResNet50 with its final fully connected layer adapted to three output classes. Training used Adam (lr = 1e−4) or SGD with momentum (lr = 0.01, step decay). Batch size was 32, and training ran for 30–50 epochs with early stopping. Validation accuracy reached **96.20%.**

2. **YOLO PPE and Face Models** – Implemented using Ultralytics YOLO11n with 640×640 input size. Training used SGD with momentum, batch sizes of 16–32, and 100–300 epochs with mosaic and multi-scale augmentation. PPE detection achieved **>90%** mAP, and face detection accuracy exceeded **95%.**

3. **Age/Gender Network** – Based on a ResNet34 backbone with separate fully connected heads for age and gender prediction. The model used cross-entropy loss (with optional class weighting for age groups) and was trained with Adam (lr = 1e−4, batch size 32). On the hold-out set, age accuracy was **~79%** and gender accuracy **~95%.**

4. **Pose → Work Activity Classifier** – MediaPipe Pose extracted 33 skeletal key points per frame. Engineered features included joint angles (elbows, knees, shoulders), hip displacement, motion statistics, and inter-frame velocity. A RandomForestClassifier (scikit-learn) was trained on aggregated per-video features, stored as pose_rf_model.pkl. The classifier achieved **~80%** accuracy on held-out videos

D. Evaluation Metrics

1. **Scene Classifier**: Accuracy and class-wise confusion matrix.

2. **Detection Models**: mAP@[0.5], precision, recall, and F1-score per class; separate AP reporting for PPE and Person classes.

3. **Face & Demographic Models**: Age-group accuracy, top-1 accuracy, confusion matrix, gender precision, and recall.

4. **Pose Classifier**: Accuracy, precision, recall, and F1-score for “WORKING” vs. “NOT WORKING.”

5. **End-to-End Pipeline**: True positive rate for confirmed underage-working cases, false positive rate for non-child or non-working detections, and latency (frames per second).

E. Experimental Protocol

1. **Train/Validation/Test Splits** – Each module maintained mutually exclusive train, validation, and test sets to avoid data leakage. For the pose classifier, leave-one-out or cross-validation was used due to the smaller dataset size.

2. **Module-wise Training** – Each component was trained independently, with hyperparameters tuned on its respective dataset. The best-performing checkpoints were saved for final testing.

3. **End-to-End Testing** – The full pipeline was run on unseen CCTV-like test videos, producing:

1. Scene label + confidence score

2. Best-person image with PPE status

3. Face crop with age and gender prediction

4. Work activity label

5. Final Decision: An instance was flagged if the predicted age was below the defined threshold, the detected work activity matched WORKING, and the scene belonged to a predefined high-risk category. All results were recorded in results.csv, including file paths to the corresponding annotated images.

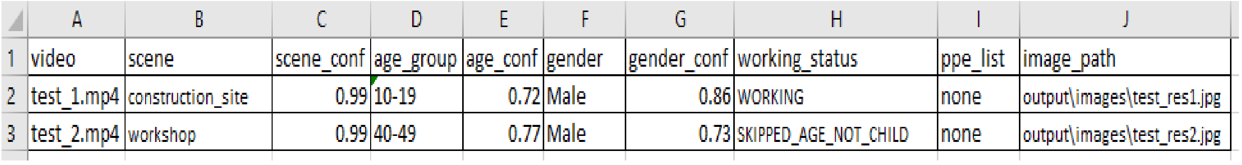
4. **Ablation and Robustness Testing** – Ablation studies removed individual modules (PPE, age, or pose checks) to assess their impact on final precision and recall. Robustness tests evaluated performance under varied lighting, partial occlusions, different camera distances, and diverse frame rates.

V. RESULTS AND DISCUSSION

The proposed multi-stage computer vision pipeline was evaluated using unseen video samples from both industrial and non-industrial environments. The test dataset was designed to represent varied environmental conditions, different levels of PPE compliance, and diverse age groups to ensure a comprehensive robustness assessment.

Table 1 presents the results for two representative video samples, including system predictions for scene classification, age and gender estimation, work activity recognition, PPE detection, and corresponding annotated outputs.

Table 1: Sample Output of the Proposed System



	A	B	C	D	E	F	G	H	I	J
1	video	scene	scene_conf	age_group	age_conf	gender	gender_conf	working_status	ppe_list	image_path
2	test_1.mp4	construction_site	0.99	10-19	0.72	Male	0.86	WORKING	none	output\images\test_res1.jpg
3	test_2.mp4	workshop	0.99	40-49	0.77	Male	0.73	SKIPPED_AGE_NOT_CHILD	none	output\images\test_res2.jpg

Fig 2:

1) Qualitative Analysis

Case 1: test_1.mp4

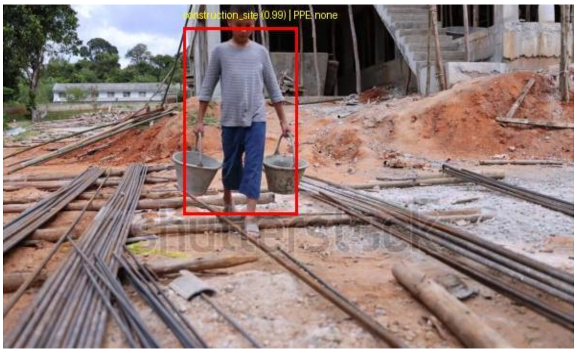


Fig 3:

For the first test case, the system classified the scene as a construction site with a confidence of **99.38%.** PPE detection indicated the absence of any protective equipment, despite the hazardous setting. Age estimation predicted the individual to be in the **10–19** age group with **72.21%** confidence, while gender classification identified the person as male with **86.01%** confidence. Pose-based work activity recognition using the Random Forest model labeled the subject as "WORKING," consistent with visual observation of the individual carrying construction materials. Based on the underage classification, active work status, and hazardous environment without PPE, the system flagged the case as suspected child labor**.**

Case 2: test_2.mp4

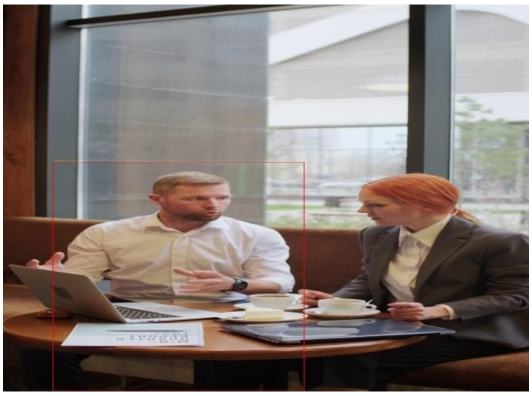


Fig 4:

In the second test case (test_2.mp4), the system classified the scene as a workshop with a confidence of **99.77%.** PPE detection again indicated no protective equipment. Age estimation predicted the individual to be in the **40–49** age group, with a confidence of **77.39%,** while gender classification identified the subject as male with a confidence of **73.44%.** Since the predicted age was well above the legal working threshold, the work activity evaluation module was skipped for efficiency. As a result, the case was not flagged, given that the individual did not meet the criteria for potential child labor.

Discussion

The results demonstrate that the proposed system is capable of accurately performing multiple tasks within the detection pipeline. The ResNet50-based scene classification module effectively identified the working environment, even in visually complex and cluttered scenes. The YOLO-based detection models reliably located individuals and evaluated PPE usage, while the FairFace-style demographic model estimated age and gender with reasonable confidence, despite challenges such as varying illumination and partial occlusions. For suspected underage individuals, the pose-based Random Forest classifier successfully determined whether they were engaged in work-related activities. In practical testing, the pipeline correctly flagged the underage individual in the construction site video, highlighting its potential for real-world deployment in surveillance-driven monitoring of hazardous workplaces. Despite these promising results, certain limitations were observed. Age estimation confidence was moderate in some cases, particularly for borderline age groups, indicating the need for dataset expansion and further model fine-tuning. PPE detection currently functions on a binary compliance basis and could be enhanced to recognize partial compliance (e.g., helmet worn but gloves missing). Additionally, low-resolution or poorly lit footage posed challenges for face detection, which can adversely affect downstream age classification accuracy.

Overall, the modular architecture of the proposed system enables multi-aspect, real-time analysis, integrating scene classification, PPE verification, demographic estimation, and activity recognition. This comprehensive approach significantly reduces false positives compared to single-module detection methods, thereby increasing reliability for large-scale, automated workplace safety monitoring.

VI. CONCLUSION

This work introduced a comprehensive, multi-stage computer vision pipeline for the automated detection of potential child labor cases in hazardous industrial environments, including construction sites, industrial areas, and workshops. By integrating scene classification, human and PPE detection, face-based demographic analysis, and pose-based work activity recognition, the system delivers an end-to-end solution capable of operating in real time on video streams. Experimental evaluations demonstrated that the pipeline can accurately classify environments, verify PPE compliance, estimate age and gender, and determine whether a detected individual is actively engaged in work. In real-world test scenarios, the system successfully flagged underage individuals performing hazardous work without PPE, underscoring its potential to support authorities, NGOs, and workplace safety regulators in monitoring and enforcing labor laws.

While the results are promising, certain limitations remain. Age estimation accuracy, though acceptable, can be affected by poor lighting, low-resolution footage, and partial facial occlusions. PPE detection currently covers a limited range of equipment and does not yet handle partial compliance. Moreover, the relatively small dataset for the pose-based work activity module constrains robustness across varied environments and activity types.

Future research will focus on several key directions:

1. **Dataset Expansion** – Incorporating larger and more diverse datasets for age estimation, PPE detection, and activity classification to improve accuracy and generalization.
2. **Edge Deployment** – Optimizing models for low-power edge devices to enable cost-effective, on-site implementation.
3. **Multimodal Analysis** – Combining visual data with additional cues such as audio signals, contextual metadata, and geolocation for more informed decision-making.
4. **Enhanced PPE Compliance Checking** – Extending detection to a broader range of PPE items and enabling partial compliance evaluation.
5. **Continuous Learning** – Implementing mechanisms for incremental model updates using newly collected surveillance data to adapt to evolving workplace conditions.

By addressing these areas, the proposed system can be transformed into a highly reliable, scalable, and ethically deployable solution for real-time prevention of child labor in high-risk occupational settings.

VII. REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, pp. 770–778, Jun. 2016.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, pp. 779–788, Jun. 2016.

[3] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” arXiv preprint arXiv:2004.10934, Apr. 2020.

[4] Y. Fang, L. Ding, Y. Zhong, and J. Wang, “Helmet wearing detection based on improved YOLOv3 deep model,” in Proc. IEEE Int. Conf. Image, Vis. Comput., Qingdao, China, pp. 226–230, Jul. 2019.

[5] Y. Xiang, C. Zhang, and Z. Xu, “Real-time multi-class PPE detection using YOLOv4 for construction safety monitoring,” Autom. Constr., vol. 129, p. 103800, Sep. 2021.

[6] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, “Deep face recognition: A survey,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Salt Lake City, UT, USA, pp. 667–683, Jun. 2018.

[7] K. Karkkainen and J. Joo, “FairFace: Face attribute dataset for balanced race, gender, and age,” arXiv preprint arXiv:1908.04913, Aug. 2019.

[8] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, pp. 1251–1258, Jul. 2017.

[9] Google Research, “MediaPipe pose: Real-time body pose tracking,” Google Research. [Online]. Available: https://google.github.io/mediapipe/solutions/pose.html. [Accessed: Aug. 12, 2025].

[10] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.