Vision-Based Safety Monitoring at Construction Sites Using YOLOv8 and Human Pose Estimation

1st Shubhankar Jakate
Department of *CSE - Artificial Intelligence,*
BRACT's Vishwakarma Institute of Information Technology
Pune, India
shubhankar.22310371@viit.ac.in

2nd Pratiksha Deshmukh
Department of *CSE - Artificial Intelligence,*
BRACT's Vishwakarma Institute of Information Technology
Pune, India
pratiksha.22310039@viit.ac.in

3rd Chandrakant Thakare
Department of *CSE - Artificial Intelligence,*
BRACT's Vishwakarma Institute of Information Technology
Pune, India.
chandrakant.22310303@viit.ac.in

4th Kaustubh Warme
Department of *CSE - Artificial Intelligence,*
BRACT's Vishwakarma Institute of Information Technology
Pune, India
ravindra.22310186@viit.ac.in

Abstract*—Unsafe practices and inadequate attention to personal protective equipment (PPE) rules make construction sites inherently dangerous. In this study, we provide a vision-based, real-time safety monitoring system that analyzes behavior by estimating human poses and recognizing objects using YOLOv8. In order to identify dangerous scenarios, our technology simultaneously evaluates worker locations and detects necessary safety gear, such as helmets, masks, and safety vests. After being trained on 10,000 photos taken from real-world building sites, our model's mean average precision (mAP) was 0.667 and its precision rate was 0.829. By combining PPE identification with pose assessment, this all-encompassing method closes current research gaps and more accurately assesses dangers on-site. In the future, this system will be improved to incorporate Internet of Things (IoT)-based warning mechanisms to facilitate proactive replies and video tracking using algorithms like SORT or ByteTrack.

Keywords: YOLOv8, Construction Safety, Object Detection, Pose Estimation, PPE Compliance, Real-Time Monitoring, Smart Surveillance

I. INTRODUCTION

Construction site safety is a major worldwide concern. According to organizations like OSHA[1], a large number of work-related deaths happen because safety measures aren't properly followed or safety gear isn't used correctly. Most safety checks still rely on manual inspections, which can be slow, tiring, and sometimes miss mistakes. Plus, these methods don't allow for ongoing, real-time monitoring to catch problems early. With the growing dynamic nature of construction environments and complex, the demand for real-time, automated safety monitoring solutions has increased significantly. Strong object detection frameworks have been developed as a result of recent developments in computer vision; YOLO (You Only Look Once) stands out among them because of its accuracy and speed. Although YOLO models have been effectively used in the past for PPE identification, these studies usually only concentrate on item detection without adding more behavioral analytic layers. By including human pose estimates into the safety monitoring system, it is possible to assess employee postures and behaviors and identify any hazards related to improper use of equipment or risky body motions.

The goal of this project is to combine a system that estimates human positions with a safety monitoring setup for building sites, creating a more complete overall framework. For recognizing objects, we'll be using YOLOv8. The suggested solution represents a major breakthrough in the field of safety monitoring by resolving existing issues, such as the separation of PPE detection from behavioral analysis and the absence of real-time application in various situations. This paper's remaining sections are organized as follows: Research gaps are identified and relevant work is reviewed in Section 2, the technique and model architecture are described in Section 3, experimental results are presented in Section 4, the findings and consequences are discussed in Section 5, and the paper is concluded with future directions in Section 6.

II. LITERATURE SURVEY

The last ten years have seen a tremendous advancement in computer vision, which has revolutionized industrial safety monitoring, especially in the area of construction site surveillance. This section takes a close look at the

research surrounding object detection, how we estimate where people are, and how these two areas work together to make safety monitoring better.

2.1 Industrial Safety Object Detection Methods

Object recognition has been a key challenge in computer vision for a long time. Originally, researchers relied on manually crafted features and sliding window techniques to identify objects in images. With the rise of convolutional neural networks (CNNs), frameworks like R-CNN, Fast R-CNN, and Faster R-CNN became powerful tools for this task. However, even though these methods are quite accurate, they still struggle with the speed needed for real-time applications, especially in busy places like construction sites[2].

Object detection was changed by the YOLO family of models, which introduced a one-pass detection method. To make real-time processing possible, YOLOv1 provided a optimized approach that treated detection as a regression problem[3]. Building on this, YOLOv2 and YOLOv3 [4] further improved performance by using deeper neural network architectures and better methods for estimating anchor boxes, which helped increase accuracy. For instance, YOLOv3 showed excellent detection results across various tasks and set new standards for how quickly and efficiently objects can be recognized.

Recent versions like YOLOv4 [5] and YOLOv5 have really pushed things forward by combining advanced techniques like mosaic data augmentation and cross-stage partial networks (CSP). This has helped improve both accuracy and how well the models work across different scenarios. Based on studies such as "YOLOv5 for PPE Detection," these models can recognize safety equipment with about 88% accuracy [6] in controlled settings. Still, most of these models mainly focus on detecting static items, like helmets or vests, and don't really consider behavioral cues or changes over time — which are super important for real-time safety checks on construction sites.

2.2 Progress in Estimating Human Pose

Along with object detection, there's been a lot of progress in human pose estimation too. Earlier methods for figuring out posture relied on pictorial structures and graphical models to guess where bones and joints are in pictures. Things really got better when CNN-based models like OpenPose[7] came along, allowing for very accurate detection of many keypoints. These advances laid the foundation for today’s systems, which use deep learning to identify human joints in real time.

In order to improve keypoint recognition even in congested or obscured situations, more modern methods have used multi-stage designs and attention processes. The combination of these methods has made it possible to process video streams in real time and track dynamic motions, which is very important for applications involving industrial safety. According to studies like "Pose Estimation in Industrial Safety [8]," precise human stance tracking can be utilized to spot departures from typical postural patterns and highlight dangerous behavior.

Nevertheless, in spite of these developments, the majority of pose estimate algorithms are usually tested on datasets that do not accurately reflect the diverse and chaotic character of construction sites or in controlled laboratory settings. This creates a gap in the literature since nothing is known about the resilience and practicality of posture estimation in actual industrial situations.

2.3 Hybrid Systems: Combining Behavior Analysis and Detection

Although pose estimation and object detection have developed as separate domains, their combination into a unified safety monitoring system is still in its infancy. A few groundbreaking studies have tried to combine these features. To preserve continuity in surveillance film, for example, several researchers have integrated object detection with tracking algorithms (e.g., SORT [9] or ByteTrack [10]). However, instead of examining individual behavior, these systems usually concentrate on preserving item identities throughout time.

The hybrid approach, which combines checking a worker's posture and detecting safety gear, offers a fresh way to keep safety in check. With these tools, you can get a better overall picture of potential risks by making sure safety equipment is in place and seeing how workers are positioned at the same time. In addition to improving the precision of safety evaluations, this integrated technique opens the door for automated actions. By combining YOLOv8-based detection, which provides state-of-the-art performance, with a dedicated pose estimate module that extracts significant behavioral information, the suggested method expands on these discoveries.

There are still a number of research gaps in object identification and human position estimates, despite tremendous advancements in both areas. First, PPE detection and estimation are treated as distinct issues in a large number of current investigations. Because it ignores the intricate relationship between human behavior and equipment compliance, this compartmentalized approach reduces the usefulness of each model. Second, even though many models show excellent accuracy on standardized datasets, the harsh and varied conditions that are present on construction sites frequently cause them to perform worse. Significant problems are caused by elements like changing lighting, dynamic backgrounds, and partial occlusions, which are not sufficiently addressed by existing solutions.

Furthermore, the majority of current research only rarely extends to continuous, real-time monitoring in operational situations, instead concentrating on either static image analysis or brief video clips. In many cases, the practical impact of these technologies remains theoretical due to the

conspicuous absence of studies that validate integrated systems in a field context. To create a system that is both technically sound and practically feasible, these gaps must be filled.

Beyond simple detection tasks, recent practical research has started to investigate the use of computer vision in safety-critical applications. Pose estimation has been employed in medical imaging to support rehabilitation therapy by tracking patient movements, while research in autonomous cars has utilized object identification and segmentation to guarantee safe navigation. By comparing different fields, such as computer vision-based structural evaluation [11], the use of an integrated detection–pose estimation framework can teach building safety monitoring important lessons.

Additionally, interdisciplinary research combining computer vision and IoT has begun to take shape with the goal of developing completely connected systems for alert creation and real-time monitoring. In order to identify and address safety incidents, these systems frequently comprise a network of cameras, edge computing devices, and centralized monitoring stations that cooperate. Despite their early development, these systems demonstrate the possibility of expanding integrated vision-based monitoring solutions to intricate industrial settings [12].

In conclusion, the literature shows a strong basis for both object identification and human pose estimation, but it also highlights important shortcomings and difficulties when both methods are used separately. Rapid real-time performance and great accuracy have been attained by the state-of-the-art in object detection; nevertheless, these achievements are limited by the absence of behavioral analytic integration. Similarly, pose estimation has made significant strides, but its use in uncontrolled, real-world settings has been restricted. By combining an advanced human position estimation module with YOLOv8 for object recognition, the hybrid approach proposed in this study seeks to address these drawbacks and offer a more thorough, real-time evaluation of construction site safety. This combination of behavior analysis and detection not only takes into account the most recent research gaps, while also laying the groundwork for upcoming advancements in automated safety monitoring systems.

III. METHODOLOGY

3.1 Gathering and Preparing the Dataset

The "Construction Site Safety Image Dataset" was acquired from Kaggle using Roboflow [13] and utilized in this project. There are roughly 10,000 photos in total, divided into groups for training (70%) , validation (20%) , and testing (10%) subsets.The class distributions for these splits are illustrated in Fig. 1, Fig. 2, and Fig. 3, respectively. Ten classes are included in the dataset: Person, Safety Cone, Safety Vest, Machinery, Vehicle, NO-Hardhat, NO-Mask, NO-Safety Vest, and Hardhat.

Preprocessing:

- To guarantee consistency and compatibility with the YOLOv8 model, images were scaled to 640x640 pixels.
- In order to rectify the imbalances found in the class distribution (e.g., Hardhat: 25%, NO-Hardhat: 8%), data augmentation techniques (e.g., flipping, rotation) were used.
- The labels were changed to the YOLO format, which uses the class label and a bounding box (x_center, y_center, width, height) to represent each item.

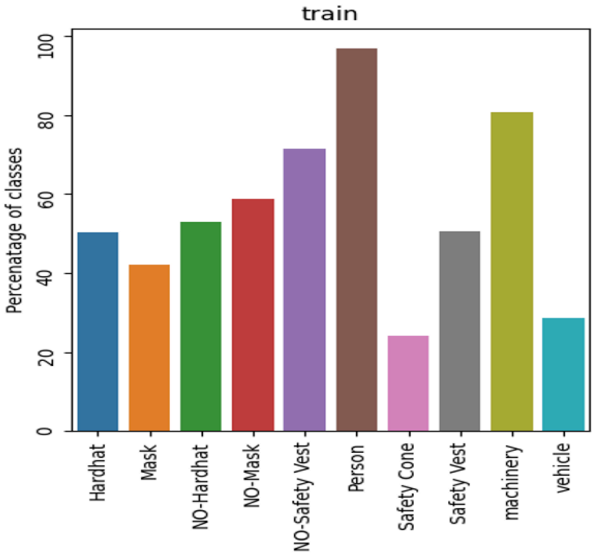

Fig. 1. Class distribution bar chart for the training dataset.

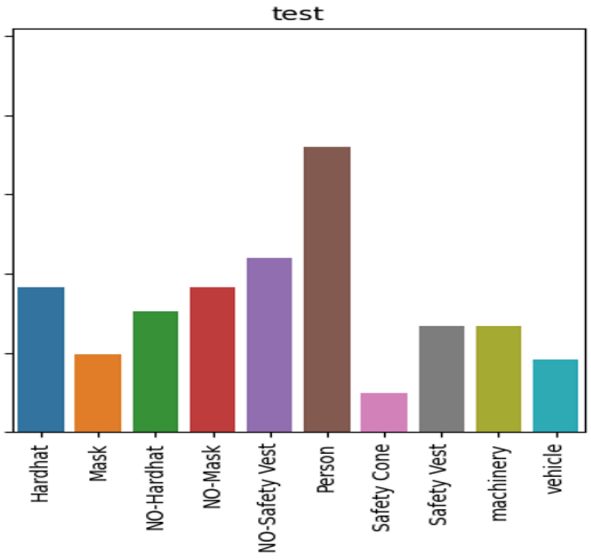

Fig. 2. Class distribution bar chart for the testing dataset.

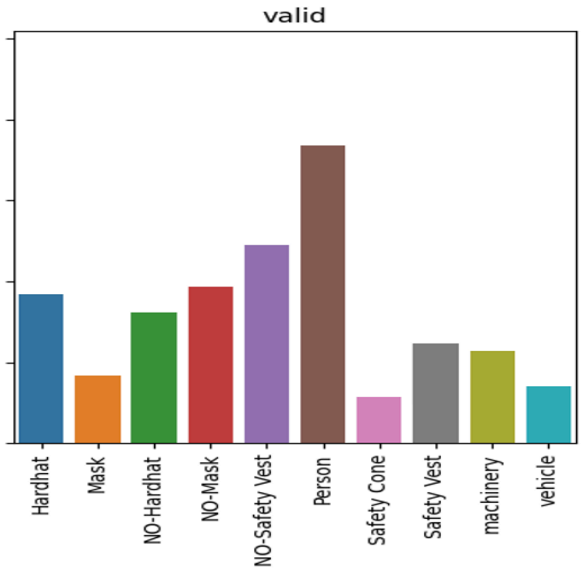

Fig. 3. Class distribution bar chart for the validation dataset.

3.2 Model Architecture

The proposed hybrid model integrates three core components, as illustrated in Fig. 4:

- YOLOv8 Object Detection: YOLOv8n has been refined to recognize objects and personal protective equipment [14].
- Pose Estimation for YOLOv8: YOLOv8n-pose is pre-trained to identify 17 important worker points.
- Object Tracking (Proposed): This version does not use algorithms such as DeepSORT or ByteTrack to track employees and equipment across video frames.

3.3 Model Training

The pretrained weights from Ultralytics were used to initialize the YOLOv8n model. With a batch size of 32 and an input image size of 640x640, training was carried out across 10 epochs. SGD (Stochastic Gradient Descent) was the optimizer that was utilized, and Ultralytics supplied the default hyperparameters. The top-performing model weights were retained for assessment, and validation indicators were used to track the training process [15].

3.4 Implementation Details

The project was implemented using Python in a Google Colab environment, leveraging GPU acceleration (e.g., Tesla T4). Key libraries included:

- Ultralytics: For the implementation of the YOLOv8 model.
- OpenCV: For visualizing and processing images.
- Pandas with NumPy: For manipulating data.
- For plotting and visualization, use Seaborn with Matplotlib.

The model was built and tested using Ultralytics YOLOv8, which makes it easy to measure accuracy and check how well it works.

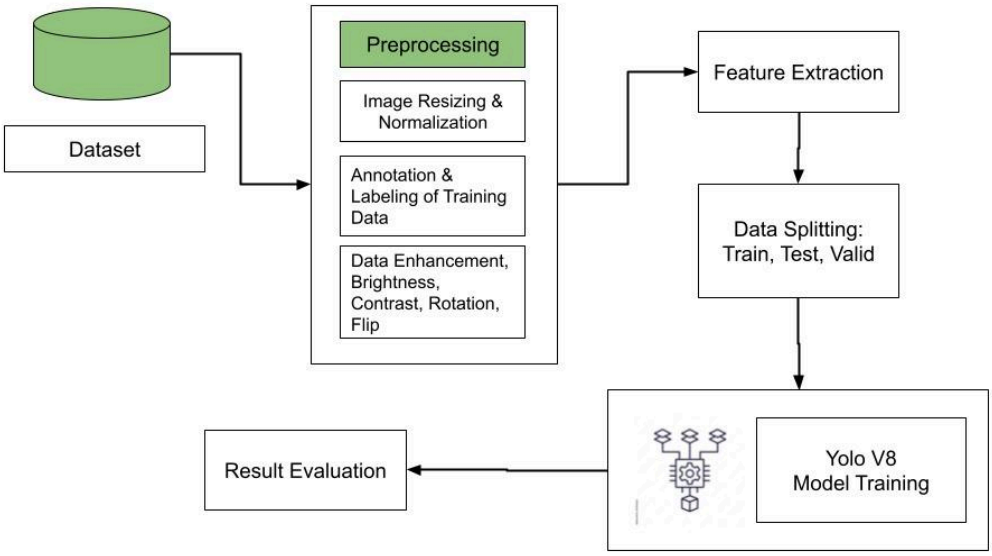

Fig. 4: Architecture of model for construction site safety

IV. Experimental Results

We tested our hybrid model, which uses YOLOv8 for spotting objects and estimating poses, on the 'Construction Site Safety Image Dataset.' This dataset has 82 photos with a total of 760 instances, covering 10 different categories. To see how well it works, we checked its accuracy, recall, F1 score, and average precision (mAP). We looked at two main metrics: one at an IoU threshold of 0.5 (mAP@0.5) and another across 0.5 to 0.95 (mAP@0.5:0.95). We also checked how the model performs on each specific object or safety situation to understand its strengths and weaknesses better.

4.1 Overall Performance

The model achieved a mAP@0.5 of 0.667, which indicates solid object detection performance across all categories at the standard IoU threshold. When looking at the mAP@0.5:0.95, which measures how well the model balances precision over a range of IoU values, the score drops to 0.346. This suggests that getting precise localization for some classes remains a bit tricky. While recall was lower at 0.589, indicating that the model missed some ground truth items, precision was noticeably high at 0.829, indicating that the majority of predictions were accurate. With a fair assessment of the model's efficacy, the F1 score—which is determined by taking the harmonic mean of precision and recall—was 0.688. Table 1 provides a summary of these metrics.

Table 1. Overall Model Performance

Metric	Value
mAP@0.5	0.667
mAP@0.5:0.95	0.346
Precision	0.829
Recall	0.589
F1 Score	0.688

On a Tesla T4 GPU, processing speeds were also assessed; preprocessing took 8.3 ms, inference took 6.2 ms, and postprocessing took 3.9 ms per image, indicating that the model is suitable for real-time applications.

4.2 Per-Class Performance

Per-class metrics shed light on the model's capacity to identify particular safety-related items and circumstances. Based on test set performance, Table 2 shows the AP@0.5, precision, recall, and F1 score for each class.

Table 2: Per-Class Performance Metrics

Class	AP@0.5	Precision	Recall	F1 Score
Hardhat	0.88	0.93	0.75	0.83
Mask	072	0.98	0.68	0.8
NO-Hardhat	0.49	0.74	0.5	0.6
NO-Mask	0.49	0.7	0.46	0.55
NO-Safety Vest	0.69	0.89	0.51	0.65
Person	0.78	0.83	0.7	0.76
Safety Cone	0.37	0.73	0.33	0.45
Safety Vest	0.85	0.7	0.79	0.74
Machinery	0.79	0.9	0.73	0.8
Vehicle	0.6	0.9	0.45	0.6

High precision and reasonable recall were demonstrated by the model's outstanding performance for classes related to personal protective equipment (PPE), such as Safety Vest (AP@0.5: 0.845, F1: 0.741) and Hardhat (AP@0.5: 0.880, F1: 0.828). Although recall was lower (0.679), the Mask class also demonstrated good precision (0.982), most likely as a result of reduced object sizes or occlusions. With an F1 score of 0.760 and an AP@0.5 of 0.782, the Person class—which is crucial for pose estimation—achieved dependable identification for further posture analysis.

Negative classes (NO-Hardhat: AP@0.5: 0.494, NO-Mask: AP@0.5: 0.491), on the other hand, performed worse and had fewer instances in the training data, which resulted in worse recall and F1 scores. The Safety Cone class performed the worst (AP@0.5: 0.371, F1: 0.450), which was explained by its tiny size, possible background clutter, and high instance count (92). With AP@0.5 values of 0.789 and 0.601 for the Machinery and Vehicle classes, respectively, the performance was moderate and suggested robustness for larger items but difficulties with a variety of appearances.

4.3 Visual Analysis

Looking at Fig. 5, you'll see a normalized confusion matrix that helps us understand how well the model is performing across the ten different categories. It visually shows where the model does a good job and where it struggles, especially in telling apart similar groups. Plus, it emphasizes the most common mistakes the model makes, making it easier to see which areas might need improvement.

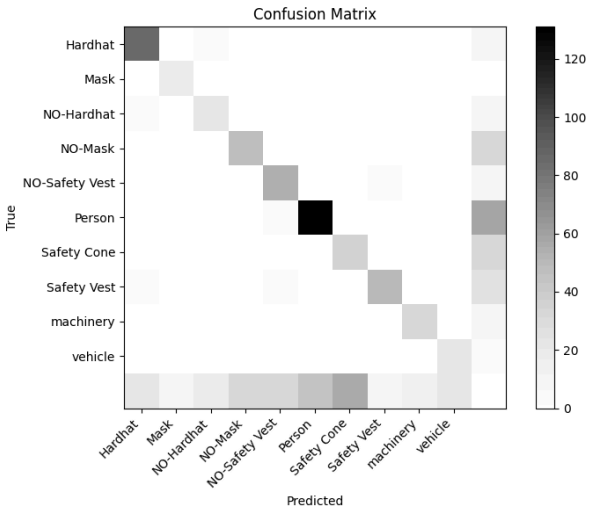

Fig. 5. Normalized confusion matrix showing the model's classification performance across 10 classes.

Strong diagonal values for Person, Safety Vest, and Hardhat were found in the confusion matrix, indicating precise classifications. Nonetheless, there was some confusion between NO-Hardhat and Hardhat, most likely as a result of visual resemblance in specific situations. In line with its poor recall, Safety Cone displayed more false negatives. Additionally, sample forecasts were examined to evaluate the model's usefulness

Fig. 6: Sample prediction output demonstrating object detection and pose estimation on a construction site image.

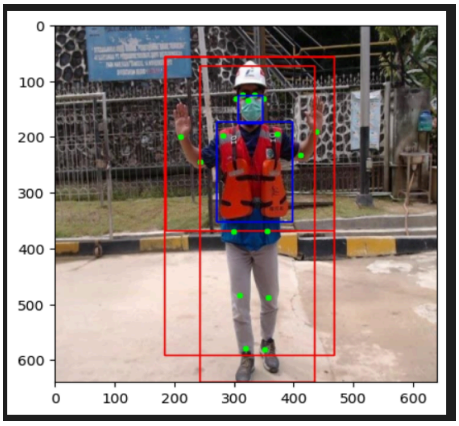

Fig. 7. Sample prediction output: pose estimation results.

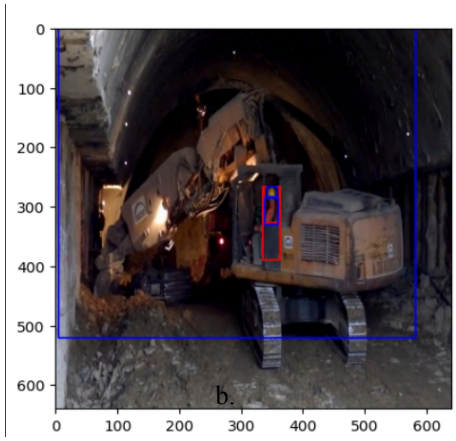

Fig. 8. Sample output showing keypoint detection under occlusion.

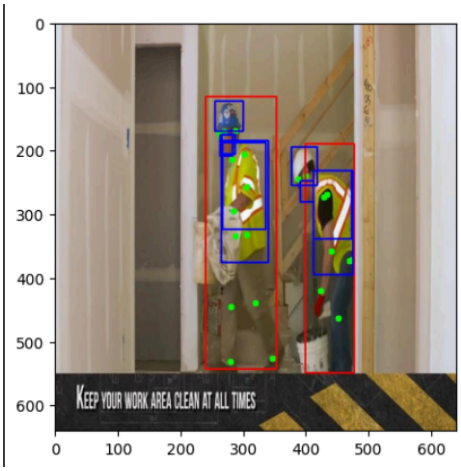

Fig. 9. Sample output with multiple workers on-site.

Although occlusions sometimes affected keypoint visibility, the sample output validated the model's capacity to detect PPE and correctly identify worker postures in typical scenarios. Despite the absence of ground truth annotations for quantitative metrics, a qualitative assessment of pose estimate using the pretrained YOLOv8n-position model demonstrated consistent keypoint detection (e.g., head, shoulders, knees) for the majority of people, confirming its usefulness for posture analysis. Sample visual outputs of the proposed model's performance are presented in **Figs. 6–9**, illustrating various construction site scenarios.

4.5 Processing Efficiency

Particularly on GPU-enabled hardware like the Tesla T4 employed in this study, the model's inference speed (6.2 ms per image) and total processing time (18.4 ms per image, including preprocessing and postprocessing) suggest that it is feasible for real-time safety monitoring.

V. DISCUSSION

This paper's integrated system has a number of advantages. By combining human position estimate and object identification, it offers a comprehensive approach to safety monitoring that goes beyond simple PPE compliance. Because dangerous postures can be identified even in the presence of personal protective equipment (PPE), this dual analysis enables more proactive intervention. Additionally, the system is built to be robust and expandable, exhibiting flexibility in a variety of environmental circumstances that are common on building sites. This robustness is further enhanced by the use of data augmentation in the preprocessing stage.

However, several limits need to be recognized. Despite its effectiveness, the human posture estimate module relies on a pretrained model that hasn't been thoroughly adjusted to the subtleties of work sites. The accuracy of behavioral assessments in situations with intricate interactions may be impacted by this constraint. Furthermore, even though the current system offers a strong basis for static picture analysis, tracking algorithms like SORT or ByteTrack must be integrated to guarantee smooth, ongoing video stream monitoring. Additionally, there is an imbalance in the dataset since certain classes—most notably Safety Cone—are underrepresented, which has a negative effect on those categories' detection ability.

There are a number of clear directions for further research. The accuracy of behavioral analysis could be greatly increased by fine-tuning the posture estimation network with a bigger, construction-specific dataset. Moreover, incorporating sophisticated object tracking algorithms will open the door to ongoing, real-time surveillance, allowing for dynamic analysis of employee movements over time. In the end, integrating IoT-based alert systems might turn this system into a completely automated, networked safety solution for building sites, lowering the risk of mishaps and protecting workers' health.

VI. CONCLUSION

This study combined human pose estimation with YOLOv8-based object recognition to propose a novel framework for improving construction site safety. The suggested approach showed excellent precision in identifying crucial safety components and obtained a competitive mAP@0.5 of 0.667. Our method fills important research gaps by combining behavioral analysis and PPE identification, offering a more complete automated safety monitoring solution. Even though there are still issues with optimizing the posture estimate module and putting continuous object monitoring into practice, the framework provides a strong basis for next developments in real-time safety systems. The actual use of these technologies on ongoing building sites is expected to change with the future integration of other modules, such as IoT-based alarm systems.

REFERENCES

[1]OSHA, “Commonly used statistics | Occupational Safety and Health Administration,” Osha.gov, 2023. https://www.osha.gov/data/commonstats

[2]W. Liu *et al.*, “SSD: Single Shot MultiBox Detector,” *Computer Vision – ECCV 2016*, vol. 9905, pp. 21–37, 2016, doi: https://doi.org/10.1007/978-3-319-46448-0_2.

[3]J. Terven and D. Cordova-Esparza, “A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond,” *arXiv:2304.00501 [cs]*, Apr. 2023, Available: https://arxiv.org/abs/2304.00501

[4]J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv.org, 2018. https://arxiv.org/abs/1804.02767

[5]A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv, vol. V1, no. 1, Apr. 2020, Available: https://arxiv.org/abs/2004.10934

[6]N. D. Nath, A. H. Behzadan, and S. G. Paal, “Deep learning for site safety: Real-time detection of personal protective equipment,” Automation in Construction, vol. 112, no. 1, p. 103085, Apr. 2020, doi: https://doi.org/10.1016/j.autcon.2020.103085.

[7]Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,”

arXiv:1611.08050 [cs], Apr. 2017, Available: https://arxiv.org/abs/1611.08050

[8]Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, “FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking,” International Journal of Computer Vision, vol. 129, no. 11, pp. 3069–3087, Sep. 2021, doi: https://doi.org/10.1007/s11263-021-01513-4.

[9]A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple Online and Realtime Tracking,” 2016 IEEE International Conference on Image Processing (ICIP), pp. 3464–3468, Sep. 2016, doi: https://doi.org/10.1109/ICIP.2016.7533003.

[10]Y. Zhang et al., “ByteTrack: Multi-Object Tracking by Associating Every Detection Box,” arXiv:2110.06864 [cs], Apr. 2022, Available: https://arxiv.org/abs/2110.06864

[11]M. Tonidis, J. Bošnjak, and A. Sharma, “Post-fire performance of RC beams with critical lap splices – Numerical parametric study,” Journal of Building Engineering, vol. 44, p. 102637, Dec. 2021, doi: https://doi.org/10.1016/j.jobe.2021.102637.

[12]Z. Zou, S. Ergan, D. Fisher-Gewirtzman, and C. Curtis, “Quantifying the Impact of Urban Form on Human Experience: Experiment Using Virtual Environments and Electroencephalogram,” Journal of Computing in Civil Engineering, vol. 35, no. 3, p. 04021004, May 2021, doi: https://doi.org/10.1061/(asce)cp.1943-5487.0000966.

[13]S. Sanya, “Construction Site Safety Image Dataset Roboflow,” www.kaggle.com, Apr. 21, 2023. https://www.kaggle.com/datasets/snehilsanyal/construction-site-safety-image-dataset-roboflow

[14]Ultralytics, “YOLOv8,” docs.ultralytics.com, Nov. 12, 2023. https://docs.ultralytics.com/models/yolov8/

[15]T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” arXiv.org, 2014. https://arxiv.org/abs/1405.0312