arXiv:2504.10880v1 [cs.CV] 15 Apr 2025

Safe-Construct: Redefining Construction Safety Violation Recognition as 3D Multi-View Engagement Task

Aviral Chharia[1], Tianyu Ren[2], Tomotake Furuhata[1], Kenji Shimada[1]
[1]Carnegie Mellon University [2]University of Illinois Urbana-Champaign
{achharia, tomotake, shimada}@andrew.cmu.edu, tianyur2@illinois.edu

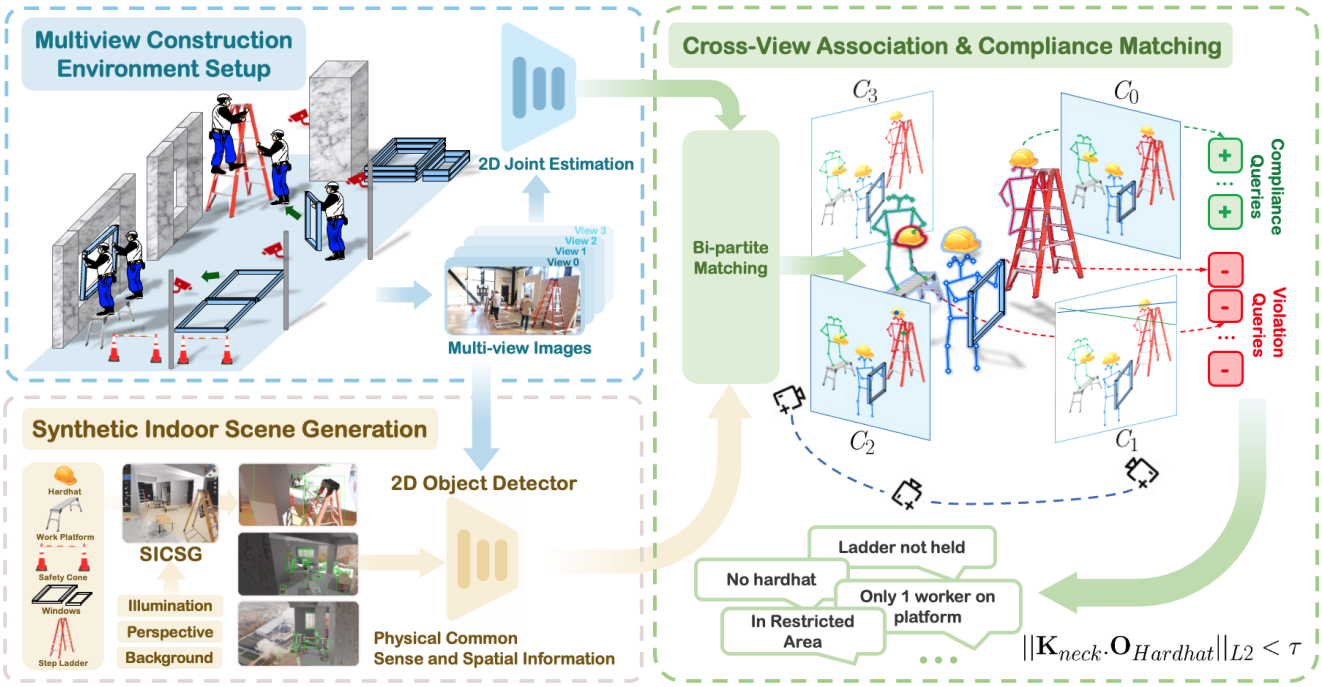


Figure 1. **Overview of Safe-Construct.** We present Safe-Construct, the first 3D multi-view safety violation recognition model for construction sites. It consists of (a) a multi-camera setup, (b) a Synthetic Indoor Construction Site Generator (SICSG), and (c) a 3D cross-view association and compliance matching module. Please check the project page for video results.

Abstract

Recognizing safety violations in construction environments is critical yet remains underexplored in computer vision. Existing models predominantly rely on 2D object detection, which fails to capture the complexities of real-world violations due to: (i) an oversimplified task formulation treating violation recognition merely as object detection, (ii) inadequate validation under realistic conditions, (iii) absence of standardized baselines, and (iv) limited scalability from the unavailability of synthetic dataset generators for diverse construction scenarios. To address these challenges, we introduce Safe-Construct, the first framework that reformulates violation recognition as a 3D multi-view engagement task, leveraging scene-level worker-object context and 3D spatial understanding. We also propose the Synthetic Indoor Construction Site Generator (SICSG) to create diverse, scalable training data, overcoming data limitations. Safe-Construct achieves a 7.6% improvement over state-of-the-art methods across four violation types. We rigorously evaluate our approach in near-realistic settings, incorporating four violations, four workers, 14 objects, and challenging conditions like occlusions (worker-object, worker-worker) and variable illumination (back-lighting, overexposure, sunlight). By integrating 3D multi-view spatial understanding and synthetic data generation, Safe-Construct sets a new benchmark for scalable and robust safety monitoring in high-risk industries. Project Website: https://Safe-Construct.github.io/Safe-Construct

1. Introduction

Construction is one of the largest sectors of the economy, generating around USD 2.2 trillion [4] and employing over

8.2 million workers in 2024–about 5% of the total U.S. workforce) [32]. Despite its scale, construction remains one of the most hazardous sectors, with 1056 fatal accidents reported in the U.S. in 2022 [33]. Nearly half of these deaths result from preventable causes, such as falls or being struck by objects [31]. These statistics highlight the importance of following safety protocols, such as wearing protective gear and adhering to guidelines for equipment like work platforms and ladders. As a result, detecting safety violations has become a critical challenge in the industry.

Vision-based methods have shown promise in improving construction safety by detecting violations. However, existing models predominantly frame this problem as an object detection task, relying on oversimplified assumptions that limit their real-world applicability [6–13, 16–19, 23, 24, 28, 29, 34–41, 45, 47–50]. Recent works have attempted to move beyond this paradigm by incorporating skeletal data [2, 46] and modeling human-object interactions [42]. Yet, these methods rely solely on 2D inputs, which often fail under heavy occlusions common at construction sites. We identify several limitations of prior works:

Lack of Realistic Training Data. Existing models [9, 28, 29, 36, 38, 45] are primarily trained on crowd-sourced or web-mined datasets, such as Pictor-v2 [28], Pictor-v3 [29], and Roboflow [38]. This results in distributions that poorly reflect real-world construction environments (see Fig. 2). They overlook perspective variations and camera distances, often assuming the camera is placed within 1 meter of the worker—an impractical setup on actual sites. Further, these datasets lack diversity in lighting conditions, such as backlighting, indoor illumination, or over-exposure, making models less robust in real-world deployment. Further, they lack appropriate 3D spatial and depth information, which is critical for recognizing violations under occlusion—such as determining whether a ladder is stabilized by a second worker when the first climbs it or identifying if large windows are being handled by multiple workers. As a result, they often focus on overly simplistic scenarios like detecting hard hats or safety vests.

Limited Scalability. Existing models rely on predefined dataset categories (*e.g.*, the presence of a "hard hat") to recognize violations. As a result, datasets must include category-level positive and negative image pairs to teach the model to identify violations as object categories. This tight coupling between violations and dataset classes limits scalability: detecting new types of violations requires collecting entirely new datasets, which is expensive, time-consuming, resource-intensive, and potentially an unsafe process.

Lack of Scene-level Understanding and Worker-Object Context. Existing approaches simplify safety violation recognition into object-centric tasks (*e.g.*, detecting a "hard hat"), neglecting broader scene-level understanding. Moreover, their training datasets typically contain isolated object images rather than full scenes, limiting the model's ability to capture contextual relationships between workers and objects, crucial for identifying intricate, real-world violations. As an example, Pictor-v3 [29] contains 1.5K images, with just 17 images including all key classes ("worker," "hard hat," and "vest"), and none showing the "worker + vest" combination. Such omissions severely constrain the development of models that necessitate scene and contextual understanding. Furthermore, they lack fine-grained annotations. For example, Pictor-v2 [28] assigns all machinery a generic "equipment" label, limiting the detection of nuanced equipment-specific violations. Such oversimplifications make current models inadequate for practical, real-world deployment.

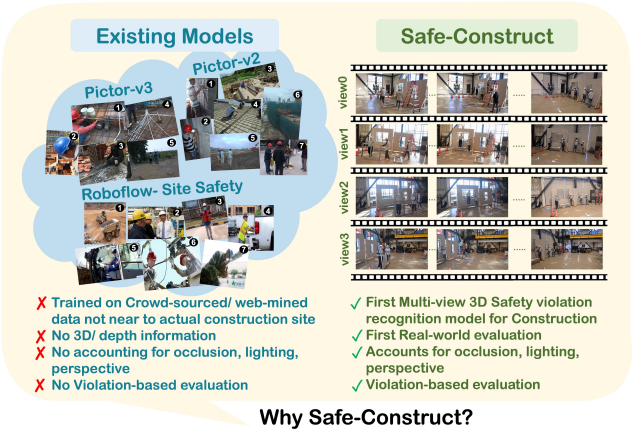


Figure 2. **Comparison with Prior Methods.** Previous models [9, 28, 29, 36, 38, 45] train on crowd-sourced or web-mined data: Pictor-v2 [28], Pictor-v3 [29], and Roboflow [38] framing the problem as 2D object detection task. These lack 3D spatial understanding and scene-level worker-object context. These datasets are small, with most images having unrealistic resolutions and perspectives (see pictures 1-4 in Pictor-v3, 1-2 in Pictor-v2, and 2-7 in Roboflow). Realistic industrial setups do not feature workers in close camera proximity (<1m). (b) In contrast, Safe-Construct is the first multi-view 3D violation recognition model that leverages 3D spatial understanding and scene-level worker-object context.

Ignoring Temporal Dynamics. Current approaches often overlook temporal consistency, which is essential for handling occlusions. Brief occlusions—such as a "hard hat" or "worker" momentarily hidden—can trigger false alarms. Additionally, when a worker's face is hidden, existing single-view models struggle to detect the person and determine if they are wearing a "hard hat".

Absence of Standardized Evaluation Metrics. Despite recent progress, construction violation recognition lacks standardized evaluation metrics. Prior studies [9, 22, 25, 26, 28, 29, 36, 38, 45] evaluate performance on diverse setups, making direct comparisons difficult. Each defines viola-

tion recognition differently, often relying on object detection metrics such as IoU-based mAP [2, 42, 46]. However, such metrics fall short in assessing a model's ability to correctly identify violations per scene, particularly in complex scenarios involving multiple objects or workers. A more meaningful metric would measure the number of violations correctly identified per scene. Unfortunately, this is not feasible with current models, which are restricted to detecting object classes like "hard hat" or "no hard hat," without capturing the violations at the full scene level. To address these gaps, we propose Safe-Construct, a novel framework for safety violation recognition. Below, we summarize our contributions:

1. We are the ***first*** to formulate violation recognition as a 3D multi-view engagement task. By leveraging geometry-based modeling and multi-view inputs, our approach achieves occlusion-robust, scene-level understanding that surpasses existing state-of-the-art methods.
2. Safe-Construct is the ***first*** framework to decouple violation criteria from training data, enabling scalable generalization to new violation types without the need for collecting additional real-dataset datasets.
3. We introduce the Synthetic Indoor Construction Scene Generator (SICSG), a novel custom engine that generates physically realistic scene variations, such as illumination, occlusion, and perspective changes, imparting spatial awareness and physical common sense to the model.
4. We conduct the ***first*** evaluation in a 3D multi-camera indoor construction setup, comprising four safety violations, four workers and 14 objects across diverse conditions—occlusions, lighting variations, and camera distances resulting in significant scale changes in worker bodies, significantly increasing scene complexity. Safe-Construct consistently outperforms prior methods. Moreover, it is the ***first*** model tailored specifically for indoor construction settings.

2. Related Works

Sensor-Based Approaches. Early efforts in construction site safety monitoring relied heavily on sensor technologies, such as RFID tags on safety equipment scanned at site entry points [30], LAN-based systems for continuous RFID tracking [1], and short-range transponders paired with wireless networks to verify workers' safety gear [21]. While effective in controlled settings, these methods demand substantial infrastructure investment and high maintenance, rendering them impractical for small construction industries.

2D Object Detection-Based Approaches. The complication and high cost of sensor networks spurred interest in vision-based methods, particularly those using affordable RGB cameras. Initial approaches [36] employed background subtraction to detect workers and assess hard hat usage but struggled with occlusions and static workers, who were often misclassified as background. Advances in RGB-based techniques have since incorporated video cues-*e.g.* motion, color information, edge detection, facial features, and Histogram of Oriented Gradients- for hard hat detection. Notable works by Nath *et al.* [28, 29] and Delhi *et al.* [9] utilized convolutional neural networks (CNNs) to identify basic safety violations. State-of-the-art efforts have predominantly adopted transfer learning with YOLO models, framing safety monitoring as an object detection task [6–8, 10–13, 16–19, 23, 24, 34, 35, 37, 39–41, 47–50]. Wu *et al.* [45] proposed a CNN tailored for hard hat recognition. However, these models often falter in generalization due to scarce construction-specific data [29] and are susceptible to occlusion-induced errors.

2D Worker Pose-Based Approaches. To address object detection limitations, recent studies have explored pose estimation for safety violation recognition. Xiong *et al.* [46] pioneered the use of 2D skeleton poses to identify construction safety breaches. Subsequent works, such as Tang *et al.*, reformulated the problem as a 2D worker-object interaction analysis, while Wang *et al.* [2] applied Graph Convolutional Networks (GCNs) for violation recognition. These methods, however, require retraining for new action classes, limiting scalability. Some approaches have also utilized depth-based cameras (*e.g.* Kinect and VICON) to analyze unsafe worker behaviors. However, their restricted field of view has shifted focus toward RGB camera solutions with broader coverage.

Leveraging Vision-Language Models. Emerging research has leveraged recent vision-language models for safety monitoring. Tsai *et al.* [43] explored fine-tuning contrastive Language-Image Pre-training (CLIP) with prefix captioning to generate automated safety observations. This direction, though promising, is nascent, with state-of-the-art models achieving accuracies of approximately 73.7%, indicating room for improvement.

3. Proposed Methodology

Unlike prior studies, we formulate violation recognition as a 3D multi-view engagement task as shown in Figure 1. Our methodology estimates 3D worker poses and construction object locations from synchronized multi-view video inputs. Leveraging multiple camera views allows us to effectively address occlusions—an inherent challenge in construction environments—while enabling a comprehensive spatial analysis of worker-object interactions from different camera perspectives.

Synthetic Scene Creation. To facilitate training of 2D object detectors for construction objects, we introduce the Synthetic Indoor Construction Site Generator (SICSG), developed in Blender 3.1.2. Prior datasets—often crowd-sourced

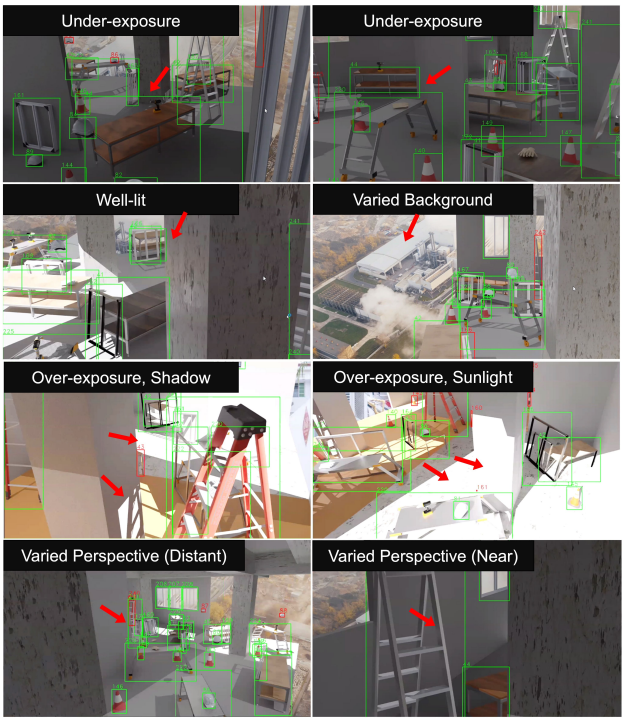


Figure 3. **SICSG Rendered Data.** Synthetic images generated by SICSG depicting variations in illumination, perspective, and backgrounds, enabling models to exhibit physical commonsense.

or web-mined—lack sufficient variability and realism for nuanced construction tasks [28, 29, 38]. SICSG overcomes these limitations by generating synthetic datasets tailored specifically to construction tasks, for example, the window installation task for our setting. It features 12 common object categories across 4000 images generated from 100 distinct room arrangements. Systematic variations in lighting, background, and camera perspective (see Fig. 3) enable models to capture spatial awareness and physical commonsense. For estimating the 2D locations of construction objects, we trained the YOLOv7 [44] architecture on dataset generated using SICSG.

2D Worker and Object Keypoint Estimation. We employ YOLOv7pose [44] to estimate 2D joint positions of workers in each camera view. Worker poses are represented in COCO format as $\mathbf{W}^{i,k} \in \mathbb{R}^{4\times17\times2}$, where i indexes the camera $(1,\ldots,N)$, and k denotes the detected worker poses per camera view $(1,\ldots,K^i)$. YOLOv7pose [44] is selected for its strong generalization ability in joint estimation tasks. Similarly, construction objects are detected individually per view as $\mathbf{O}^{i,l}$, where l indexes detected objects $(1,\ldots,L_i)$. To ensure the reliability of keypoint estimations, detections with confidence scores below a threshold ϕ are discarded.

Cross-View Association via Epipolar Geometry. Before triangulating into 3D coordinates, the detections from multi-

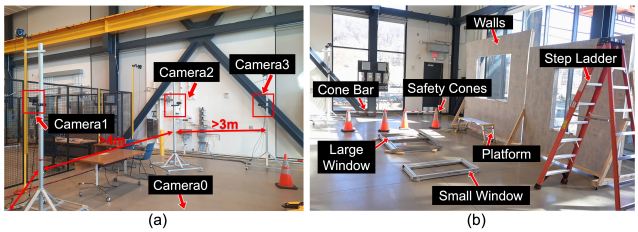


Figure 4. **Experimental setup.** The evaluation setup consisted of four time-synchronized and calibrated Nikon® FHD cameras attached to tripods. (a) Cameras 1-3 with Camera 0 out-of-view, (b) Site setting captured by Camera 0.

ple camera views must be accurately associated. For this, we employ epipolar geometry-guided bipartite matching. Initial 2D detections from the first camera view (i=1) serve as the initial worker and object selections: $\mathbf{H}_s = [\mathbf{H}^{1,k}]$ and $\mathbf{O}_s = [\mathbf{O}^{1,l}]$. Subsequently, detections from views i=2, ..., N are greedily matched with existing worker and object selections using Bi-partite matching. The assignment cost between a new detection $(\mathbf{W}^{i,k}, \mathbf{O}^{i,l})$ and existing selections $(\mathbf{W}_s^m, \mathbf{O}_s^m)$ is computed as follows:

$$\mathcal{C}_{worker}(\mathbf{W}^{i,k}, \mathbf{W}_s^m) = \frac{1}{|\mathbf{W}_c^m||J_{kl}|} \sum_{\mathbf{W}^{j,l}\in\mathbf{W}_s^m} \sum_{\mathbf{W}^{j,l}\in\mathbf{J}_{kl}} \mathcal{D}(\mathbf{W}^{i,k}(\varphi), \mathbf{W}^{j,l}(\varphi)) \quad (1)$$

$$\mathcal{C}_{object}(\mathbf{O}^{i,k}, \mathbf{O}_s^m) = \frac{1}{|\mathbf{O}_s^m||J_{kl}|} \sum_{\mathbf{O}^{j,l}\in\mathbf{O}_s^m} \sum_{\mathbf{O}^{j,l}\in\mathbf{J}_{kl}} \mathcal{D}(\mathbf{O}^{i,k}(\varphi), \mathbf{O}^{j,l}(\varphi)) \quad (2)$$

where $\mathbf{W}^{i,k}(\varphi)$ denotes the 2D-pixel location of joint φ of the 2D worker $\mathbf{W}^{i,k}$ and J_{kl} is the set of joints that are visible and non-occluded for both poses $\mathbf{W}^{i,k}$ and $\mathbf{W}^{j,l}$. The same is extended to objects with the number of joints always being one. The distance between two joints in the respective cameras is defined by the distance between the epi-polar lines and the joint locations:

$$\mathcal{D}(p_i, p_j) = |p_j^T F^{i,j} p_i| + |p_i^T F^{j,i} p_j| \quad (3)$$

where $F^{i,j}$ is the fundamental matrix from Camera i to Camera j. Solving the bipartite matching problem

$$X^* = \min_X \sum_{m=1}^{|\mathbf{W}_s^m|} \sum_{k=1}^{K_i} C(\mathbf{W}^{i,k}, \mathbf{W}_s^m) X_{k,m} \quad (4)$$

$$\sum_k X_{k,m} = 1 \;\forall\; m \text{ and } \sum_m X_{k,m} = 1 \;\forall\; k \quad (5)$$

$X^*_{k,m} = 1$ if $\mathbf{W}^{i,k}$ is associated with an existing worker selection, else 0. If $X^*_{k,m} = 1$ and $\mathcal{C}_{worker}(\mathbf{W}^{i,k}, \mathbf{W}_s^m) < \theta$, a match is confirmed and the 2D detection $\mathbf{W}^{i,k}$ is added



Figure 5. **Illumination Variations.** Sample lighting scenarios encountered during testing of Safe-Construct. The first row shows well-lit indoor rooms common in indoor construction; the second row presents backlighting and over-exposure, which reduces visibility and presents a darker appearance.

to $\mathbf{W}_s^m$. Otherwise, the detection is treated as a new entity and $\mathbf{W}^{i,k}$ is added as selection for a new worker to $\mathbf{W}$. This process is repeated for objects.

3D Triangulation. Finally, we triangulate the associated detections using calibrated camera parameters. For construction objects, we compute a weak pose by triangulating the bounding box center, suitable for small items (*e.g.*, hard hat) occupying $<1\%$ of image pixels (see Fig. 7). Despite minimal pixel coverage, this method proves reliable for 3D localization, as small objects can be approximated as points in 3D space.

Bipartite Matching for Tracking. To track the entities over time, we employ bipartite matching [27]. The newly triangulated 3D worker and weak-object pose at time t are matched to the previous 3D poses at time t-1 by bipartite matching. The matching cost is computed as the Euclidean distance between all corresponding joints that are present in both poses. If this cost is below a predefined threshold (denoted as ψ), the poses are considered matched; otherwise, they are treated as new workers or objects. In cases where two consecutive poses lack joints or have missing joints, the mean of all present joints is calculated. When computing the cost under such circumstances, the points are projected onto the x-y plane before measuring the Euclidean distance between them.

Compliance and Violation Queries. The resulting 3D worker and object pose facilitate straightforward safety compliance verification through simple geometric queries based on L2-norm. For *e.g.*, four typical safety violations are identified using thresholds derived from worker body dimensions as shown below:

- No hard hat:

$$||\mathbf{W}_{neck} - \mathbf{O}_{hardhat}||_{L2} < \tau_1 \quad (6)$$

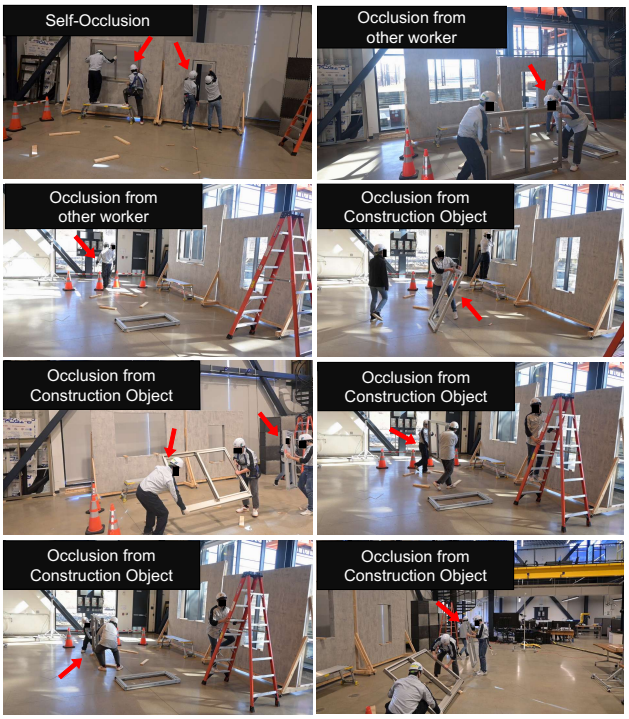


Figure 6. **Occlusions Variations.** Examples of self-occlusion and secondary occlusion during Safe-Construct testing.

- Single-worker platform restriction:

$$||\mathbf{W}_{feet} - \mathbf{O}_{platform}||_{L2} < \tau_2 \quad (7)$$

- Ladder holding requirement:

$$||\mathbf{W}^{(1)}_{torso} - \mathbf{O}_{ladder}||_{L2} < \tau_3, ||\mathbf{W}^{(2)}_{torso} - \mathbf{O}_{ladder}||_{L2} < \tau_3 \quad (8)$$

- Large window requires dual-worker handling:

$$||\mathbf{W}^{(1)}_{torso} - \mathbf{O}_{window}||_{L2} < \tau_4, ||\mathbf{W}^{(2)}_{torso} - \mathbf{O}_{window}||_{L2} < \tau_4 \quad (9)$$

Thresholds $\tau_1, \ldots, \tau_4$ are empirically derived and scaled relative to standard worker body dimensions. For example, τ_1 is set as one-tenth of the worker height W_{height}. Since the camera coordinate system normalizes scale, these thresholds are effective and invariant across different worker body sizes.

4. Experiments and Results

4.1. Experimental Setup

In this section, we present the experimental setup, results, and ablation studies for the safety violation recognition task in construction environments.

Camera Setup and Synchronization. The experimental configuration involved up to four workers, 14 objects, and four safety violations, captured by four synchronized and

Methodology	Dataset	Testing Set Specifications				Violation Recognition	
		Collection Method	Type	Scene-Level	Evaluation Metric	Classes	#Violations†
Object Detection-based Formulation							
BGR Subtraction	Park *et al.* [36]	Site	Video	-	Confusion Matrix	Hardhat, Worker	1
CNN	GDUT-HWD [45]	Web	Image	✗	IoU mAP	Hardhat, No hardhat	1
YOLOv3 Transfer Learning	Delhi *et al.* [9]	Site, Web	Image	-	Confusion Matrix	4 Safety Items	2
YOLOv2-v3 Transfer Learning	Pictor-v2 [28]	Crowd, Web	Image	✗	IoU mAP	3 Site Elements	-
YOLOv3 Transfer Learning	Pictor-v3 [29]	Crowd, Web	Image	✗	IoU mAP, Confusion Matrix	3 PPE Types	2
YOLOv8 Transfer Learning	Roboflow [38]	Crowd, Web	Image	✗	IoU mAP, Confusion Matrix	10 Site Elements	2
2D Human Pose/ Object Interaction							
Detection-based HOI	Tang *et al.* [42]	Crowd, Web	Image	-	IoU mAP	22 Categories	4
Part-aware Localization	CPPE [46]	Web	Image	✗	IoU mAP	7 PPE States	3
GCN	Wang *et al.* [2]	Site	Video	-	Confusion Matrix	6 Behaviors	5
3D Multi-view Engagement Task Formulation							
Safe-Construct (Ours)	SICSG	Synced Setup	Video	✓	Scene-level Confusion Matrix	4 Dynamic Cases	4‡

Table 1. **Comparison with Prior Methods.** Safe-Construct compared to existing models (we do not consider RGBD). †± Categories of same violation counted once ‡ Evaluated on four cases. Safe-Construct violations are not dataset-dependent.

calibrated Nikon FHD RGB cameras (indexed 0-3), each with a resolution of 1920×1080 at 30 FPS. To ensure stability, all cameras were mounted on tripods. The target area was designed to emulate a real-world construction site by maximizing its size while maintaining a converging layout. As illustrated in Fig. 4, the cameras were positioned in a semi-circle around the target area at a height of 2.0 meters. Distances from the cameras to the scene center ranged from 3-10 meters, enabling the capture of large-scale scenes with significant variations in worker scale. Synchronization was achieved via a wired interface and connected to a central server, with time offsets computed between each camera and the server's clock to ensure precise temporal alignment, following prior motion capture studies [5, 15, 20]. Timestamp-based file naming was followed to store temporal data.

Camera parameters were estimated using checkerboards following OpenCV standards [3, 51], with Camera 0 as the reference origin. Due to the large target area, checkerboards were not simultaneously visible across all four cameras, necessitating a pairwise calibration approach for extrinsics. Intrinsic calibration yielded low RMSE values of 0.2567, 0.1823, 0.2092, and 0.1623 (pixels) for Cameras 0-3, and an extrinsic RMSE of 0.8098, 0.5806, and 1.1369 for Cameras 1-3 relative to Camera 0.

Illumination and Occlusion Variations. Real-world construction sites are characterized by diverse lighting conditions and frequent occlusions. Figures 5 and 6 illustrate the test setup, which represents these challenges. Lighting variations included indoor illumination, back-lighting scenarios with sunlight exposure, and over-exposure conditions. Occlusions ranged from partial or complete obstruction of workers by construction objects to self-occlusions caused by intricate worker movements or poses during tasks.

Safety Violation Scenarios. We evaluated four intricate and complex real-world safety violations specific to indoor window installation. This included: (1) No hard hat, (2) Step Ladder not stabilized/ held by a second worker when the first worker climbs it, (3) Carrying a large window requires dual handling, (4) Only one worker at a time is allowed on a platform. Unlike previous models, scene-level annotations were applied consistently across all frames of a sequence, even if a violation was occluded in some views. For instance, if a safety violation occurred, all four frames were labeled with that violation as ground truth, even if the violation was not visible due to occlusion in one of the frames. The model was implemented on an NVIDIA RTX-3090 GPU with 64GB RAM.

Why Accurate Pose Estimation is Not the Focus? We do not compare our model's results with multiview 3D human pose estimation benchmarks, such as CMU-Panoptic [20] or Human3.6M [14], as these are primarily intended for 3D human pose studies without containing construction objects or safety violation recognition. Moreover, pose distortions do not critically impair violation recognition (see Figure 9) as geometric triangulation suffices for our use case. In our use case, real-time pose estimation is prioritized over achieving an extremely accurate pose because timely recognition is critical for practical applications in construction environments. Therefore, geometric triangulation was employed without additional methods to enhance pose accuracy, which is out of the scope of the current study.

4.2. Results and Discussion

We test our model for four violations in the challenging over-exposure case. For comparison with previous approaches, we re-trained the current SOTA model on the same synthetic dataset from SICSG. Herein, we designed a 2D single-frame violation recognition model using the object-detection-based approach. For a fair comparison, we used the same backbone model and L2-norm-based safety criteria and thresholds. Further, to make the comparison even more challenging,

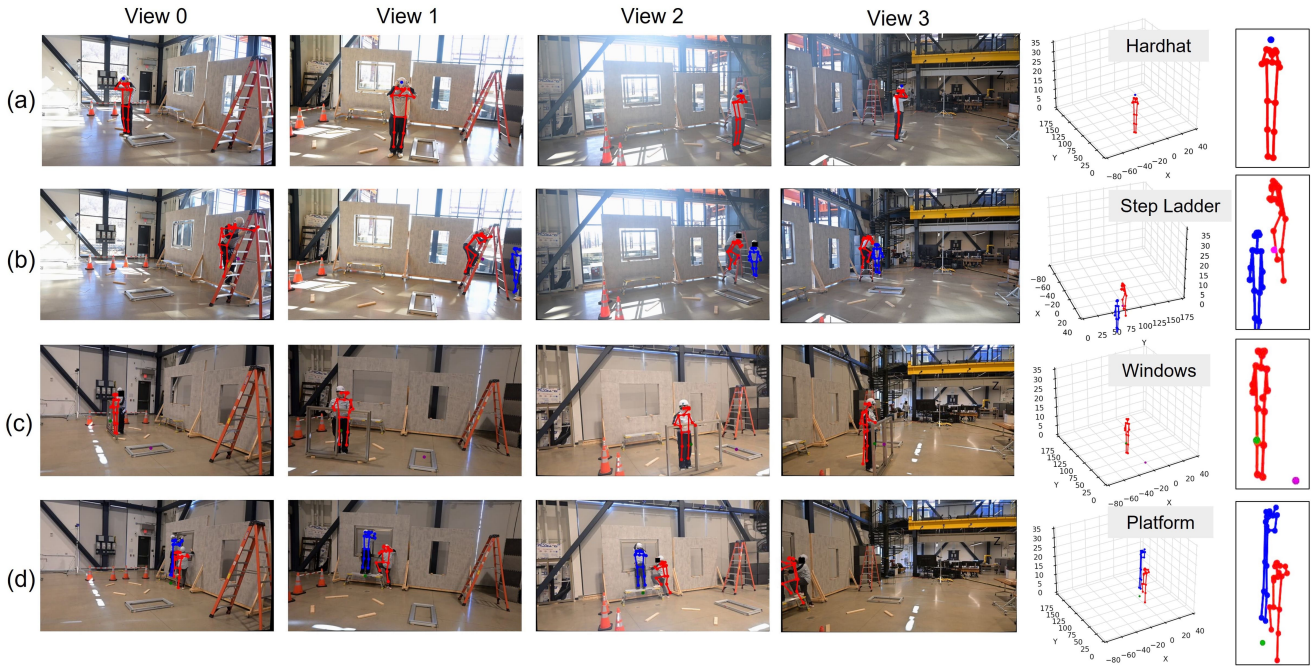


Figure 7. **Safety Violation Recognition Qualitative Results.** We show the 2D re-projection of worker and object's pose on the image plane: Rows (a), (b) show two safe scenarios, while (c), (d) illustrate two violations: (a) Worker wearing a hard hat, (b) Second worker holding the Step ladder when the first climbs it, (c) Only one worker is carrying a Large window that should be carried by two workers showing a violation scenario (the small window is represented in magenta), (d) Two workers are standing on the Platform simultaneously.

Models	Supervision	Safety Violation Recognition Accuracy (%)				AVG.
		No Hard hat	Ladder not stabilized by the second worker when the first climbs	Large Windows requires handling by two workers	Only one worker at a time is allowed on the platform	
Baseline	One view	81.7	78.2	84.5	91.4	-
Ours	Two views	84.0	81.6	87.2	91.8	**+2.2%**
	Improvement (abs)	**+2.3%**	**+3.4%**	**+2.7%**	**+0.4%**	
	Four views	92.0	88.4	91.4	94.2	**+7.6%**
	Improvement (abs)	**+10.3%**	**+10.2%**	**+6.9%**	**+2.8%**	

Table 2. **Quantitative Comparison.** Performance for the violation recognition task. Values are rounded to the first decimal place.

we utilized camera i=1, which offered the front-view and is the most advantageous. This is also consistent with industrial setups, wherein the front camera is the main input. Table 2 reports recognition accuracy across the four violation types. The baseline achieves 81.7–91.4% accuracy, while our method with two views improves performance by an average of +2.2%, and with four views by +7.6%. Notable gains include +10.3% for "No hard hat" and +10.2% for "Ladder Not Held", highlighting the advantage of multi-view supervision. In summary, our multi-view system significantly outperforms single-view baselines, with a four-view setup achieving up to +10.3% accuracy gains. These results validate the efficacy of multi-camera supervision for safety violation recognition in complex, occluded environments. We present the qualitative results in Figures 7, 8, and 9, including limitations and failure cases.

Figure 7 illustrates 2D re-projections of worker and object poses, with rows (a)–(b) depicting safe scenarios (*i.e.*, worker with hard hat, ladder held) and rows (c)–(d) showing violations (*i.e.*, single worker with large window, crowded platform). Edge cases (see Figure 8), such as missed hard hat recognition due to occlusion, are mitigated by temporal consistency and multi-view data.

Effect of Multiple-view 3D Spatial Understanding. To assess the impact of multiple views, compared to single-camera setups, model performance was evaluated across varying supervision (Table 2). Two views enhance violation recognition over the baseline (*e.g.*, +3.4% for "Ladder Not Held"), while four views yield the highest gains, particularly in occlusion-heavy scenarios. Results indicate that increasing the number of views from one to four improves average



Figure 8. **Edge Cases.** (a) A case when the hard hat is not detected. We mark this as a violation, based on the previous frame *i.e.*, unless the worker wears the hard hat again, all frames are tagged as violations. (b) Increasing the number of views improves model prediction.

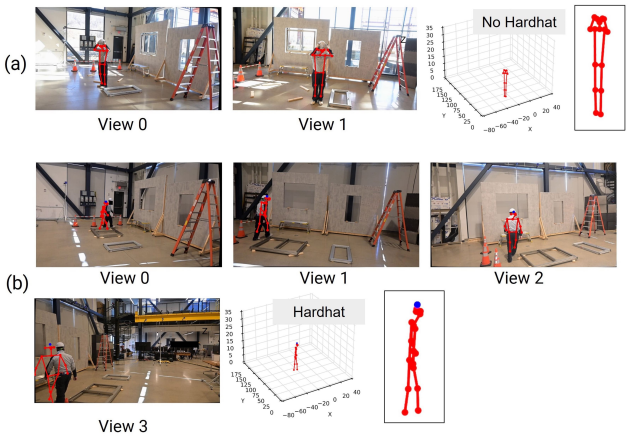


Figure 9. **Failure Cases.** (a) Results using 2-views highlighting missed hard hat prediction. (b) Pose distortions.

performance by $+7.6\%$. This underscores the robustness of a multi-view setup for addressing occlusions.

5. Conclusion and Future Work

We present Safe-Construct, a novel 3D violation recognition model as a step towards addressing the limitations of current models in multi-view indoor construction environments. Experimental results demonstrate Safe-Construct's superiority to previous methods. We believe that Safe-Construct will catalyze the development of algorithms designed to sense violations on construction sites, leading to improvement in safety in these industrial environments.

One constraint of Safe-Construct lies in its reliance on precise camera calibration. To emulate the real world, our setup consisted of a large target area. We utilized pairwise stereo calibration to calibrate the cameras. However, this has a compounding effect on RMSE values across multiple pairwise calculations. If not calibrated precisely, this can adversely affect the model's performance. Additionally, motion blur—stemming from the 30 FPS capture rate—poses a challenge, as high-frame-rate cameras are uncommon in real-world construction settings. Thus, developing algorithms resilient to motion blur is critical. Other avenues for future work include: *(i)* analyzing the trade-off between violation recognition accuracy and precision of 3D joint estimation for workers, and *(ii)* optimizing runtime performance by retraining a pose estimation backbone that focuses solely on the worker joints that are relevant to the violation criteria.

Acknowledgments. The work was supported by YKK-AP Inc., Japan. Aviral Chharia was supported in part by the ATK-Nick G. Vlahakis Graduate Fellowship from Carnegie Mellon University, Pittsburgh. The authors are thankful to Aman Chulawala, Hidetaka Kajikawa, Naoto Tanaka, and Yoriko Ogawa for helpful discussions and feedback to improve the work. Work was done while Tianyu Ren was at Carnegie Mellon University.

References

[1] Santiago Barro-Torres, Tiago M Fernández-Caramés, Héctor J Pérez-Iglesias, and Carlos J Escudero. Real-time personal protective equipment monitoring system. *Computer Communications*, 36(1):42–50, 2012. 3

[2] Wang Bo, Ma Fuqi, Jia Rong, Luo Peng, and Dong Xuzhu. Skeleton-based violation action recognition method for safety supervision in the operation field of distribution network based on graph convolutional network. *CSEE Journal of Power and Energy Systems*, 2021. 2, 3, 6

[3] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 6

[4] United States Census Bureau. Construction spending. https://www.census.gov/construction/c30/prpdf.html, 2024. 1

[5] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October*

23–27, 2022, Proceedings, Part VII, pages 557–577. Springer, 2022. 6

[6] Rong Chang, Bingzhen Zhang, Qianxin Zhu, Shan Zhao, Kai Yan, and Yang Yang. Ffa-yolov7: Improved yolov7 based on feature fusion and attention mechanism for wearing violation detection in substation construction safety. *Journal of Electrical and Computer Engineering*, 2023(1):9772652, 2023. 2, 3

[7] BinBin Chen, XiuHui Wang, GuanCheng Huang, and GuangYan Li. Detection of violations in construction site based on yolo algorithm. In *2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, pages 251–255. IEEE, 2021.

[8] Liang Cheng. A highly robust helmet detection algorithm based on yolo v8 and transformer. *IEEE Access*, 2024. 3

[9] Venkata Santosh Kumar Delhi, R Sankarlal, and Albert Thomas. Detection of personal protective equipment (ppe) compliance on construction site using computer vision based deep learning techniques. *Frontiers in Built Environment*, 6: 136, 2020. 2, 3, 6

[10] Dong Ding, Zhengrong Deng, and Rui Yang. Yolo-tc: An optimized detection model for monitoring safety-critical small objects in tower crane operations. *Algorithms*, 18(1):27, 2025. 3

[11] Ikchul Eum, Jaejun Kim, Seunghyeon Wang, and Juhyung Kim. Heavy equipment detection on construction sites using you only look once (yolo-version 10) with transformer architectures. *Applied Sciences*, 15(5):2320, 2025.

[12] Ahatsham Hayat and Fernando Morgado-Dias. Deep learning-based automatic safety helmet detection system for construction safety. *Applied Sciences*, 12(16):8268, 2022.

[13] Hoseong Hwang, Donghyun Kim, and Hochul Kim. Fd-yolo: A yolo network optimized for fall detection. *Applied Sciences*, 15(1):453, 2025. 2, 3

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 6

[16] Md Shariful Islam, SM Shaqib, Shahriar Sultan Ramit, Shahrun Akter Khushbu, Abdus Sattar, and Sheak Rashed Haider Noori. A deep learning approach to detect complete safety equipment for construction workers based on yolov7. *arXiv preprint arXiv:2406.07707*, 2024. 2, 3

[17] Xuejun Jia, Xiaoxiong Zhou, Zhihan Shi, Qi Xu, and Guangming Zhang. Geoiou-sea-yolo: An advanced model for detecting unsafe behaviors on construction sites. *Sensors*, 25(4): 1238, 2025.

[18] Xin Jiao, Cheng Li, Xin Zhang, Jian Fan, Zhenwei Cai, Zhenglong Zhou, and Ying Wang. Detection method for safety helmet wearing on construction sites based on uav images and yolov8. *Buildings*, 15(3):354, 2025.

[19] Peijian Jin, Hang Li, Weilong Yan, and Jinrong Xu. Yolo-esca: a high-performance safety helmet standard wearing behavior detection model based on improved yolov5. *IEEE Access*, 12: 23854–23868, 2024. 2, 3

[20] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):190–204, 2019. 6

[21] Agnes Kelm, Lars Laußat, Anica Meins-Becker, Daniel Platz, Mohammad J Khazaee, Aaron M Costin, Manfred Helmus, and Jochen Teizer. Mobile passive radio frequency identification (rfid) portal for automated and rapid control of personal protective equipment (ppe) on construction sites. *Automation in construction*, 36:38–52, 2013. 3

[22] Hongjo Kim, Hyoungkwan Kim, Yong Won Hong, and Hyeran Byun. Detecting construction equipment using a region-based fully convolutional network and transfer learning. *Journal of computing in Civil Engineering*, 32(2):04017082, 2018. 2

[23] Yuanfeng Lian, Jing Li, Shaohua Dong, and Xingtao Li. Hr-yolo: A multi-branch network model for helmet detection combined with high-resolution network and yolov5. *Electronics*, 13(12):2271, 2024. 2, 3

[24] Yongqiang Liu, Pengxiang Wang, and Haomin Li. An improved yolov5s-based algorithm for unsafe behavior detection of construction workers in construction scenarios. *Applied Sciences*, 15(4):1853, 2025. 2, 3

[25] Xiaochun Luo, Heng Li, Dongping Cao, Fei Dai, JoonOh Seo, SangHyun Lee, et al. Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. *J. Comput. Civ. Eng*, 32(3):04018012, 2018. 2

[26] Milad Memarzadeh, Mani Golparvar-Fard, and Juan Carlos Niebles. Automated 2d detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automation in Construction*, 32:24–37, 2013. 2

[27] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957. 5

[28] Nipun D Nath and Amir H Behzadan. Deep convolutional networks for construction object detection under different visual conditions. *Frontiers in Built Environment*, 6:97, 2020. 2, 3, 4, 6

[29] Nipun D Nath, Amir H Behzadan, and Stephanie G Paal. Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, 112: 103085, 2020. 2, 3, 4, 6

[30] Berardo Naticchia, Massimo Vaccarini, and Alessandro Carbonari. A monitoring system for real-time interference control on large construction sites. *Automation in Construction*, 29: 148–160, 2013. 3

[31] U.S. Department of Labor. Occupational safety and health administration commonly used statistics. https://www.osha.gov/data/commonstats, 2019. 2

[32] U.S. Bureau of Labor Statistics. Industries at a glance: construction. https://www.bls.gov/iag/tgs/iag23.htm, 2023. 2

[33] U.S. Bureau of Labor Statistics. Census of fatal occupational injuries summary. https://www.bls.gov/news.release/cfoi.nr0.htm, 2023. 2

[34] Oğuzhan Önal and Emre Dandıl. Unsafe-net: Yolo v4 and convlstm based computer vision system for real-time detection of unsafe behaviours in workplace. *Multimedia Tools and Applications*, pages 1–27, 2024. 2, 3

[35] Minsoo Park, Jinyeong Bak, Seunghee Park, et al. Small and overlapping worker detection at construction sites. *Automation in Construction*, 151:104856, 2023. 3

[36] Man-Woo Park, Nehad Elsafty, and Zhenhua Zhu. Hardhat-wearing detection for enhancing on-site safety of construction workers. *Journal of Construction Engineering and Management*, 141(9):04015024, 2015. 2, 3, 6

[37] Hirald Dwaraka Praveena, Goruva Vardhan Babu, Jinka Muni Sai Yateesh, Degasikshi Guna, and Guduru Vamsi Krishna. Automated helmet detection and fine issuance system using yolov8 & opencv. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 2241–2246. IEEE, 2025. 3

[38] Roboflow Universe Projects. Construction site safety dataset. https://universe.roboflow.com/roboflow-universe-projects/construction-site-safety, 2023. visited on 2024-03-04. 2, 4, 6

[39] Yash Seth and M Sivagami. Enhanced yolov8 object detection model for construction worker safety using image transformations. *IEEE Access*, 2025. 3

[40] Nihal Pattan Shetty, Jolepalyam Himakar, Pathi Gnanchandan, Suresh Jamadagni, et al. Enhancing construction site safety: A tripartite analysis of safety violations. In *2024 3rd International Conference for Innovation in Technology (INOCON)*, pages 1–6. IEEE, 2024.

[41] R Sudharshan, N Kandavel, S Latheesh Saran, et al. Automated dress code compliance monitoring using yolo based models. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 1973–1978. IEEE, 2025. 2, 3

[42] Shuai Tang, Dominic Roberts, and Mani Golparvar-Fard. Human-object interaction recognition for automatic construction site safety inspection. *Automation in Construction*, 120: 103356, 2020. 2, 3, 6

[43] Wei-Lun Tsai, Phuong-Linh Le, Wang-Fat Ho, Nai-Wen Chi, Jacob J Lin, Shuai Tang, and Shang-Hsien Hsieh. Construction safety inspection with contrastive language-image pre-training (clip) image captioning and attention. *Automation in Construction*, 169:105863, 2025. 3

[44] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 4

[45] Jixiu Wu, Nian Cai, Wenjie Chen, Huiheng Wang, and Guotian Wang. Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. *Automation in Construction*, 106:102894, 2019. 2, 3, 6

[46] Ruoxin Xiong and Pingbo Tang. Pose guided anchoring for detecting proper use of personal protective equipment. *Automation in Construction*, 130:103828, 2021. 2, 3, 6

[47] Kai Yan, Quanjing Li, Hao Li, Haifeng Wang, Yuping Fang, Lin Xing, Yang Yang, Haicheng Bai, and Chengjiang Zhou. Deep learning-based substation remote construction management and ai automatic violation detection system. *IET generation, transmission & distribution*, 16(9):1714–1726, 2022. 2, 3

[48] Guoliang Yang, Xinfang Hong, Yangyang Sheng, and Liuyan Sun. Yolo-helmet: A novel algorithm for detecting dense small safety helmets in construction scenes. *IEEE Access*, 2024.

[49] Juntao Zan, Yang Fang, Qilie Liu, Uswah Khairuddin, Yan Li, and Kaiwei Sun. Mkd-yolo: Multi-scale and knowledge-distilling yolo for efficient ppe compliance detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[50] Yi Zhang, Shize Huang, Jinzhe Qin, Xingying Li, Zhaoxin Zhang, Qianhui Fan, and Qunyao Tan. Detection of helmet use among construction workers via helmet-head region matching and state tracking. *Automation in Construction*, 171:105987, 2025. 2, 3

[51] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 6