Context-Adaptive CCTV Pan-Tilt-Zoom method for Personal Protective Equipment Detection

Seokhwan Kim[1], Minwoo Jeong[1], Minkyu Koo[1], Taegeon Kim[1], and Hongjo Kim[1]

[1]Department of Civil and Environmental Engineering, Yonsei University, Seoul, South Korea

yyksh2019@yonsei.ac.kr, minuj9855@yonsei.ac.kr, kmk0119804@yonsei.ac.kr, ktg9655@yonsei.ac.kr, hongjo@yonsei.ac.kr (Corresponding Author)

Abstract -

PPE items, including hardhats, hooks, harnesses, and straps, are critical for fall prevention. Ongoing research in construction safety has focused on using deep learning models to detect Personal Protective Equipment (PPE) worn by high-altitude workers. Despite efforts using computer vision-based models for safety monitoring, small object detection, such as hooks and straps, remains challenging due to image resolution issues. This study introduces a novel technique using mobile CCTV cameras controlled by an automated Pan-Tilt-Zoom (PTZ) algorithm to enhance the detection of small-sized PPE. The method leverages the size gap between worker and PPE. In a zoomed-out state with a short focal length, the system identifies the worker's bounding box (b-box), then zooms in with a longer focal length for precise PPE detection. When encountering multiple workers, the system applies predetermined zoom-in rules. Experimental results demonstrated a significant increase in detection accuracy for the small PPE: hook detection improved from 39.8% to 88.3%, and strap detection from 49.4% to 71.8%, as measured by an mAP of 50. This encouraging performance improvement suggests that automated PTZ control technology could enhance the effectiveness of safety monitoring

Keywords -

Construction safety; PTZ CCTV control; monitoring; PPE detection; Small object detection

1 Introduction

Construction industry, notorious for the highest number of accident victims, experiences most fatalities due to falls, according to the U.S. Bureau of Labor Statistics [1]. In South Korea, the most common type of industrial accidents is also falls, leading to a significant number of injuries [2]. Falls typically occur from high workspaces due to slipping or structural failures. Despite attempts to install safety nets and personal airbags, their high costs and spatial constraints limit widespread usage in construction sites. Therefore, proper attachment and usage of Safety Hooks and Safety Straps to fixed structures such as scaffolding are crucial to prevent falling from height position.

To increase the use of PPE, construction companies globally, including in South Korea, are employing construction site safety managers to monitor workers' compliance with PPE usage, with some regions enforcing this as a legal requirement. However, this manual method is labor-intensive, costly, and prone to human error. Recent advancements have seen the integration of CCTV and cutting-edge deep learning for automated safety monitoring in construction sites. These systems use deep learning-based object detection or instance segmentation models to identify workers and PPE from video frame, automatically determining proper PPE usage. This automated safety monitoring tries to aid or replace human managers, reducing errors, cutting costs, and increasing efficiency in construction projects.

Figure 1. Video captured by CCTV installed at an actual construction site.

While automated safety monitoring technology has been continually developing, achieving significant success in certain areas such as recognizing workers' behaviors, detecting hardhats, or differentiating between high-altitude and ground-level workers, it still faces substantial challenges in recognizing small PPE items such as safety hooks and straps. This difficulty arises because these PPE items are often captured as only a few to several pixels in size, making it hard for object recognition models to identify their features. For example, Figure 1 shows this challenge. It is easy to recognize five workers in ground and high-altitude, but hard to find their hooks and straps,

even if human eyes.

Figure 2. Results of applying Super-Resolution (SR) to construction site image. (a,b) show before SR application, and (c,d) are after applying SR. (b) and (d) specifically provide enlarged views of a worker in the top right area.

Several software-based solutions have been proposed to address low-resolution issues in images. For example, Super Resolution (SR) technology, aimed at enhancing image resolution software-wise, has been evolving for decades [3][4] and now focuses on deep learning-based methods [5]. Despite its advancements, SR's application in identifying small PPE at construction sites often doesn't meet expectations. Figure 2 shows the use of a contemporary deep learning-based SR method [6] on construction videos. Overall image clarity improves from (a) before SR to (c) after SR application. However, there are limitations when comparing (b) and (d), where despite clearer worker visibility, finding small PPE remains problematic. Thus, the approach of only using post-processing on already captured video frame has limitations in effectively detecting small PPE.

This study tackles the issue of small object detection, such as PPE detection in far-field monitoring. The goal is to overcome problems with low-resolution images by first taking larger pictures of PPE. The study leverages PTZ CCTV cameras with remote control capabilities, employing a worker-centric approach for zooming in to fill video frames, thus enlarging small PPE for better visibility. This research is potentially pioneering in proposing an automated system for adjusting PTZ cameras specifically for PPE detection in construction workers. It utilizes the size different between a typical 180cm worker and 20cm PPE, prioritizing worker detection in a zoomed-out state, then calculating center coordinates for targeted zoom-in, enhancing focused on the PPE detection. To evaluate the effectiveness of the suggested system, experiments were carried out in a laboratory environment, capturing both zoomed-in and zoomed-out videos targeting a worker. Performance of the PPE detection model was compared between two scenarios. The feasibility of implementing automatic PTZ control was examined in an indoor setting. The integration of automated PTZ CCTV control with safety monitoring in this study is expected to demonstrate utility and facilitate precise safety monitoring. Moreover, the system's design to operate without the need for personnel to control the CCTV leads to a more efficient monitoring system, reduces human error in surveillance, and is anticipated to improve the accuracy of monitoring.

2 Related works

2.1 Construction safety monitoring with computer vision

In recent years, continuous research has utilized advanced computer vision technology for safety monitoring in construction sites [7]. Studies include training Region-based fully convolutional networks to recognize construction equipment [8], installing CCTV on cranes for worker safety monitoring in complex construction environments [9], and combining computer vision models with IoT sensors for accurate fall hazard detection [10]. Additionally, depth estimation on single-lens captured images has been proposed for improved safety monitoring [11], and optimizing loss functions in deep learning models has been shown to enhance PPE detection [12]. These studies indicate widespread use of computer vision in construction safety monitoring and support the suitability of instance segmentation and object detection for PPE detection. However, they primarily focus on recognition performance in already captured footage, with less discussion on the methods of capturing the footage itself.

2.2 Autonomous PTZ control

Automating PTZ control has been extensively researched. Maximizing PTZ CCTV's object tracking and zooming capabilities is key. Attempts include using classical computer vision methods such as KLT feature trackers for PTZ operation [13]. Efforts to reduce or mitigate delays between video and PTZ control have been made for successful zooming and tracking [14]. Studies on optimally operating multiple PTZ cameras in a space, considering field of view and concurrent tracking, have been conducted [15]. Research on inspection robots for continuous object monitoring [16] and integrating PTZ control with neural networks for end-to-end solutions [17] have also been explored. Previous research on PTZ cameras

typically involved wired connections, enabling detection without significant delays, which differs from this study. Here, the CCTV is wirelessly connected, resulting in a delay of about 3 seconds for video reception and motor control. This delay was factored into the design of the automated PTZ control system. The interval between PTZ control commands had to be longer than this delay to prevent malfunctioning.

3 Methodology

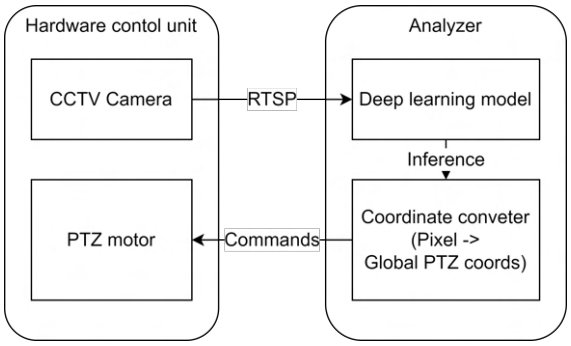

Figure 3. Simplified flowchart of the automated PTZ CCTV system

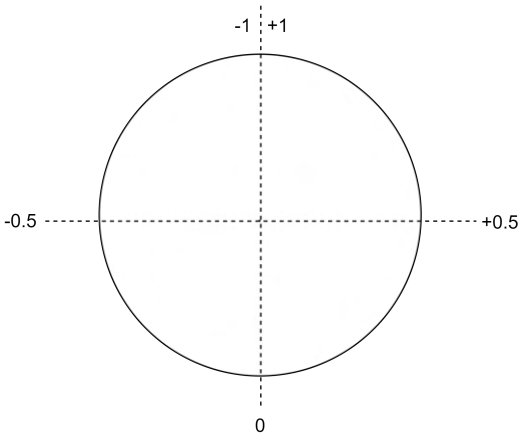

Figure 4. Top view of the PTZ CCTV's PAN coordinates

Figure 3 illustrates the automated PTZ CCTV system, which consists of a hardware control unit for operating the PTZ and an analyzer unit that processes footage from the CCTV. The CCTV continuously captures frames and broadcasts it using the Real-Time Streaming Protocol (RTSP), a standard protocol commonly used in CCTV systems for transmitting live video. The analyzer receives this video via RTSP for analysis. Deep learning models em-

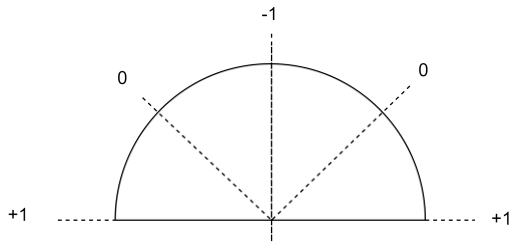

Figure 5. Side view of the PTZ CCTV's TILT coordinates

ployed for analysis include object detection and instance segmentation models such as YOLOv8m-seg [18]. These models are trained to identify workers and PPE, with their training process described later in 3.2. The system calculates the actual coordinates on the CCTV camera from the pixel coordinates of objects detected by the model Figure 4 and Figure 5.

$$\phi_{target} = \frac{x - \frac{W}{2}}{\frac{W}{2}} * \frac{FOV_w}{2} + \phi_{center} \quad (1)$$

$$\theta_{target} = \frac{y - \frac{H}{2}}{\frac{H}{2}} * \frac{FOV_h}{2} + \theta_{center} \quad (2)$$

The detailed method for coordinate calculation follows equations 1 and 2. W and H represent the width and height pixel values of the image, also x and y mean the pixel coordinates of target object in the image. FOV_w and FOV_h refer to the Field of View (FOV) of the camera in the width direction and height direction. ϕ_{target} and θ_{target} represent the Pan and tilt coordinates of target object. ϕ_{center} and θ_{center} represent the center coordinates of image.

Based on the estimated coordinate information, the PTZ motor is activated through a Open Network Video Interface Forum(ONVIF) protocol which is a global standard for the interface of IP-based physical security products, such as network cameras. This processes are repeated, continuously capturing and analyzing footage and operating the PTZ control. The performance of the proposed system is validated by the improved mask-AP(Average Precision) of the deep learning model.

3.1 States and transitions

A Finite State Machine (FSM) is a design method where a device can only exist in one of a finite number of states at a time [19]. The FSM allows a system to operate within predictable states, enabling stable control. The proposed system's PTZ control algorithm is designed as an FSM, as shown in Figure 6. The diamond in the figure represents the starting state, while the circles represent other

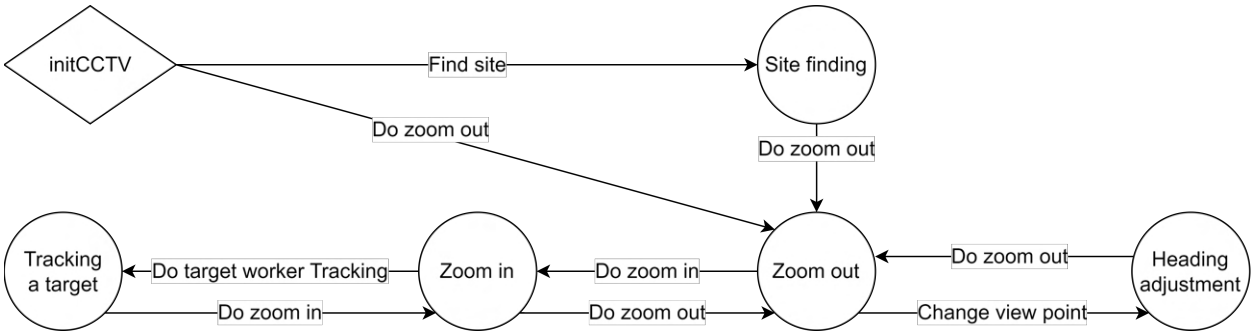

Figure 6. Flow chart of proposed auto PTZ control system using finite state machine

states. The lines between states indicate possible transitions. The system transitions from one state to another based on predefined procedures. Details on actions and transition definitions for each state will be discussed in the subsequent sections.

3.1.1 initCCTV

The 'initCCTV' state activates when the system starts. It has no inward transition since it's the initial state. In this state, the system checks the CCTV power, initiates the ONVIF protocol, positions Pan-Tilt-Zoom to zero, and creates shared memory between states. These operations prepare the system for action, initialize hardware, and stabilize the system. It also checks communication status, attempting reconnection if issues arise. Transition to the next state is based on predefined user instructions: it transitions to 'Find site' if commanded, or 'Do zoom out' otherwise.

3.1.2 Site finding

The 'Site Finding' state enables a PTZ CCTV to automatically detect and orient towards the direction of ongoing construction work, allowing it to start filming independently without remote assistance via human. Figure 7 illustrates how this feature operates. Initially, it performs 'Heading to zero position,' returning pan and tilt to positions 0 and 1. To anticipate network delays, a refresh function clears any buffer backlog. Then, it captures a single frame from the CCTV. 'n' represents the number of captures, dividing 360º by 'n' to determine the pan angle per capture. The deep learning model identifies workers in each frame, storing their locations. This process repeats until a full rotation is completed. Afterwards, the number of detected workers at each rotation point is averaged to identify the current work site, and the camera is oriented accordingly. Once complete, the system transitions to 'Zoom out' state.

3.1.3 Zoom out

In the 'Zoom out' state, the system receives RTSP and searches for workers or, if possible, PPE in the video. This state manages the overall schedule, alternating every 5 seconds to the 'Heading adjustment' state or switching to the 'Zoom in' state every 30 seconds.

3.1.4 Heading adjustment

The 'Heading adjustment' state involves receiving real-time RTSP video to locate workers and automatically adjust the camera's direction towards them. Once this action is completed, the system reverts back to the 'Zoom out' state.

3.1.5 Zoom in

The 'Zoom in' state uses the PTZ's zoom feature to select a worker for closer observation. After detecting a worker and calculating their PTZ coordinates, it zooms in on a certain worker based on a pre-chosen policy: (1) smallest area worker, (2) no PPE, or (3) from left to right sides. The extent of zooming is until the worker's bounding box (b-box) fills the video frame. If the setting for 'Tracking a target' is enabled, it activates the 'Tracking a target' state every 3 seconds. The 'Zoom in' state operates for 15 seconds, after which it returns to the 'Zoom out' state.

3.1.6 Tracking a target

In the 'Tracking a target' state, the system continuously follows the magnified individual. It calculates the necessary pan and tilt adjustments considering the focal length changes due to zooming. This process lasts for 5 seconds, after which the system reverts back to the 'Zoom in' state.

3.2 PPE detection model

Accurately and rapidly detecting workers and PPE is crucial for effective safety monitoring. For this, YOLOv8,

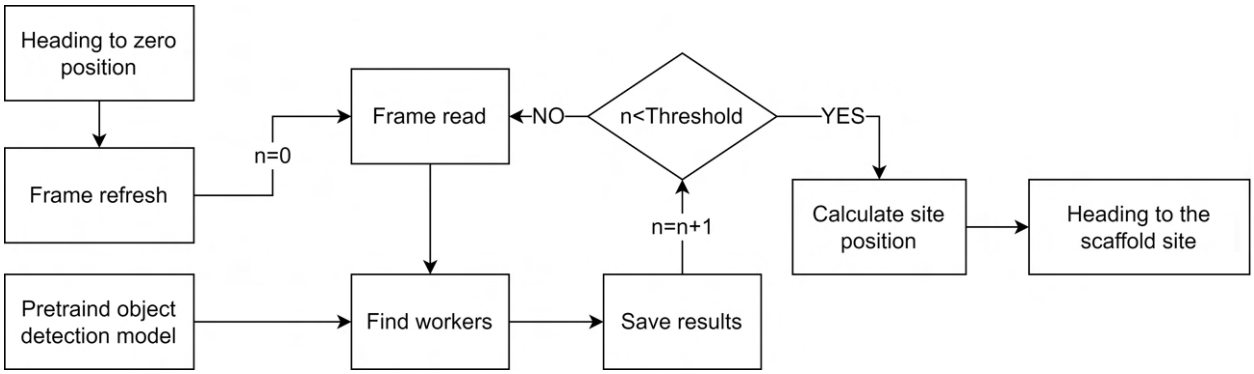

Figure 7. Algorithm of the Site Finding

known for its accuracy and detection speed, was employed. YOLOv8 introduced an anchor-free detection system, enhancing performance with faster computation and better accuracy. Mosaic augmentation, used until 10 epochs before training completion, prevented overfitting, ensuring general detection capabilities. YOLOv8's versatility allows easy modification or addition of features to its head, if needed. Thus, the study adopted YOLOv8m-seg, using transfer learning on data labeled in instance segmentation format from a construction site video collected during 2022-2023 in Korea.

4 Experiment

4.1 Experimental Settings

Figure 8. The PTZ CCTV is mounted on a module equipped with a router and battery.

The PTZ CCTV camera used in this study is the Hikvision 'DS-2DE4A225IW-DE 2MP' model, capable of up to 25x zoom and providing a 57.6º FOV at 1x zoom shown in Figure 8. The RTSP video streams at a resolution of 1280x720 at 10 frames per second (FPS). The computer used for training and inference employs an RTX3090 GPU.

4.2 Image dataset for PPE detection model

Figure 9. Example: (Left) Images and (Right) GT masks

The dataset used for training the model consists of videos collected from 65 construction sites, including apartment and road projects in South Korea, labeled for instance segmentation with four classes: worker, hardhat, strap, and hook (as shown in Figure 9). It comprises a total of 6,523 images, divided into training, validation, and testing subsets in a ratio of 5,877:600:46 for use in training.

4.3 Train the model

The training of the YOLOv8m-seg model followed the default settings suggested in [18], with the only modification being an increase in the maximum epochs to 300. This model pretrained on the MS COCO dataset, and utilized transfer learning in this research to develop a fast and high-performing model. The dataset used was the one introduced in 4.2, focusing on learning and locating features of workers, hardhats, straps, and hooks.

4.4 Evaluation matric

The evaluation metric used was the mask_AP. It assesses instance segmentation by calculating the Intersection over Union (IoU) between the predicted mask and the

true mask, considering instances with IoU over 50% as True Positives (TP). This method of evaluating instance segmentation performance centered on masks incorporates both recall and precision of predictions, offering a comprehensive evaluation of the segmentation's accuracy.

4.5 Lab test of developted PTZ control

Figure 10. (Left) Zoom-out view and (Right) Zoom-in view

Video data of workers and PPE was collected from temporary structures at Yonsei University. This data comprises 230 images, simulating zoomed-in and zoomed-out states as shown in Figure 10. The images were polygon labeled for the same four classes as mentioned in 4.2. This dataset is utilized to assess the practical effectiveness of the zoom-in.

5 Results and Discussion

5.1 Performance of the PPE instance segmentation model

The training results on the dataset from 4.2 showed outcomes as in Table1. High mask_mAP performances of 97.1% for 'Worker' and 95.2% for 'Hardhat' were achieved, whereas 'Strap' and 'Hook' exhibited lower performances at 60% and 48.2%, respectively. This reaffirms the difficulty in recognizing small PPE in far field situations.

Table 1. Performance of the model

Class	mask_mAP@50
Worker	97.1
Hardhat	95.2
Strap	60.0
Hook	48.2

The developed model was applied to the small-scale dataset of zoomed-in and zoomed-out images in 4.5, and its performance was evaluated. As seen in Table2, significant performance improvements were noted for 'Strap' and 'Hook'. In zoomed-out situations, 'Strap' and 'Hook' showed lower performances of 49.4% and 39.8%, respectively, while zoomed-in, they exhibited remarkable improvements with 71.8% and 88.3%.

Table 2. Performance of the model between zoom-in and zoom-out

Class	mask_mAP@50	
	Zoom-out	Zoom-in
Worker	99.5	99.5
Hardhat	99.5	99.5
Strap	49.4	71.8
Hook	39.8	88.3

5.2 Qualitative results of automated PTZ system

It was observed that according to the pre-determined rules of the finite state machine, the transition from the Zoom-out state to Zoom-in and Tracking a target occurs as illustrated in Figure 11. With each activation of the zoom, making small PPE more detectable by eye becomes apparent. In contrast, without PTZ control, detection is limited to workers or hardhats only.

Figure 11. Comparison of situations with and without automated PTZ peration

Applying these experimental findings to real construction sites could greatly improve the accuracy of detecting whether PPE is worn, thereby significantly boosting site safety. Moreover, the capability to automatically track workers from a distance allows for effective monitoring of the site, regardless of CCTV camera placement.

6 Conclusion

This study presented a comprehensive examination of the implementation and efficacy of an automated PTZ CCTV system for enhancing safety monitoring on construction sites. Our research found that zooming in on small objects such as hooks or straps significantly enhances detection capabilities. Laboratory experiments

with zoomed-in and zoomed-out footage, analyzed using the same model, indicated a substantial improvement in performance—by approximately 1.6 to 2 times. Moreover, the zoom-in and tracking states proved effective in detecting small PPE items previously undetectable.

Still, there are limitations to be addressed for further advancement of the proposed PTZ control method:

1. **Need for Field Data Validation:** Testing in real construction environments is necessary to validate the PTZ control system's efficacy, thereby revealing unknown issues that hamper the reliable monitoring system.

2. **Improvement in Coordinate Calculation at Lower Tilt Values:** Future study should focus on enhancing the accuracy of coordinate calculations, especially at lower Tilt angles which are the error sources reducing the current system's precision.

3. **Delays in Video Transmission Over Wireless Networks:** Future study should explore the integration of edge computing and the PTZ CCTV. These efforts are intended to reduce the data transmission time, improving the system's responsiveness.

4. **Efficient Zoom-In Target Selection:** Additionally, identifying a systematic and efficient method for determining zoom-in targets will be essential. This will ensure the PTZ control system can focus on relevant areas quickly and accurately, enhancing its utility in monitoring safety equipment on construction sites.

By addressing these issues and possible solutions, future study aims to significantly improve the PTZ control method's reliability and effectiveness.

The research contributes valuable insights into the field of construction safety and lays the groundwork for future innovations that could potentially automate and improve safety measures, thereby reducing the risk of accidents and enhancing worker protection in construction environments.

Acknowledgment

This research was conducted by the support of the "2023 Yonsei University Future-Leading Research Initiative (No. 2023-22-0114)" and the "National R&D Project for Smart Construction Technology (No. RS-2020-KA156488)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation.

References

[1] U.S. BUREAU OF LABOR STATISTICS. Construction deaths due to falls, slips, and trips increased 5.9 percent in 2021. On-line: https://www.bls.gov/opub/ted/2023/construction-deaths-due-to-falls-slips-and-trips-increased-5-9-percent-in-2021.htm, Accessed: 25/12/2023.

[2] Ministry of Employment and Labor. 2022 industrial accident statistics: Results of 'investigation of fatal accidents' released. On-line: https://www.moel.go.kr/news/enews/report/enewsView.do?news_seq=14546, Accessed: 25/12/2023.

[3] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3): 21–36, 2003. doi:10.1109/MSP.2003.1203207.

[4] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 349–356, 2009. doi:10.1109/ICCV.2009.5459271.

[5] Dawa Chyophel Lepcha, Bhawna Goyal, Ayush Dogra, and Vishal Goyal. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91:230–260, 2023. doi:https://doi.org/10.1016/j.inffus.2022.10.007.

[6] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 36081–36093, 2022. doi:https://doi.org/10.48550/arXiv.2207.08494.

[7] Weili Fang, Lieyun Ding, Peter ED Love, Hanbin Luo, Heng Li, Feniosky Pena-Mora, Botao Zhong, and Cheng Zhou. Computer vision applications in construction safety assurance. *Automation in Construction*, 110:103013, 2020. doi:https://doi.org/10.1016/j.autcon.2019.103013.

[8] 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. doi:https://doi.org/10.1061/(ASCE)CP.1943-5487.0000731.

[9] Gelayol Golcarenarenji, Ignacio Martinez-Alpiste, Qi Wang, and Jose Maria Alcaraz-Calero. Machine-learning-based top-view safety monitoring of ground workforce on complex industrial sites. *Neural Computing and Applications*, pages 1–14, 2022. doi:https://doi.org/10.1007/s00521-021-06489-3.

[10] Muhammad Khan, Rabia Khalid, Sharjeel Anjum,

Si Van-Tien Tran, and Chansik Park. Fall prevention from scaffolding using computer vision and iot-based monitoring. *Journal of Construction Engineering and Management*, 148(7):04022051, 2022. doi:https://doi.org/10.1061/(ASCE)CO.1943-7862.0002278.

[11] Wei-Chih Chern, Jeongho Hyeon, Tam V Nguyen, Vijayan K Asari, and Hongjo Kim. Context-aware safety assessment system for far-field monitoring. *Automation in Construction*, 149:104779, 2023. doi:https://doi.org/10.1016/j.autcon.2023.104779.

[12] Wei-Chih Chern, Tam V Nguyen, Vijayan K Asari, and Hongjo Kim. Impact of loss functions on semantic segmentation in far-field monitoring. *Computer-Aided Civil and Infrastructure Engineering*, 38(3):372–390, 2023. doi:https://doi.org/10.1111/mice.12832.

[13] Keni Bernardin, Florian van de Camp, and Rainer Stiefelhagen. Automatic person detection and tracking using fuzzy controlled active cameras. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. doi:10.1109/CVPR.2007.383502.

[14] Gengjie Chen, Pierre-Luc St-Charles, Wassim Bouachir, Guillaume-Alexandre Bilodeau, and Robert Bergevin. Reproducible evaluation of pan-tilt-zoom tracking. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2055–2059, 2015. doi:10.1109/ICIP.2015.7351162.

[15] Samer Hanoun, James Zhang, Vu Le, Burhan Khan, Michael Johnstone, Michael Fielding, Asim Bhatti, Doug Creighton, and Saeid Nahavandi. A framework for designing active pan-tilt-zoom (ptz) camera networks for surveillance applications. In *2017 Annual IEEE International Systems Conference (SysCon)*, pages 1–6, 2017. doi:10.1109/SYSCON.2017.7934744.

[16] Yong Li, Liang Pan, and Tao Cheng. A camera ptz control algorithm for autonomous mobile inspection robot. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 962–967, 2021. doi:10.1109/ICBAIE52039.2021.9389970.

[17] Sandeep Singh Sandha, Bharathan Balaji, Luis Garcia, and Mani Srivastava. Eagle: End-to-end deep reinforcement learning based autonomous control of ptz cameras. *arXiv preprint arXiv:2304.04356*, 2023. doi:https://doi.org/10.48550/arXiv.2304.04356.

[18] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics. On-line: `https://github.com/ultralytics/ultralytics`, Accessed: 25/12/2023.

[19] Ferdinand Wagner, Ruedi Schmuki, Thomas Wagner, and Peter Wolstenholme. *Modeling software with finite state machines: a practical approach*, volume 1. CRC Press, 2006.