Automation in Construction 130 (2021) 103828

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Pose guided anchoring for detecting proper use of personal protective equipment

Ruoxin Xiong, Pingbo Tang *

Dept. of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ARTICLE INFO

Keywords:
Personal protective equipment (PPE)
Pose estimation
Part attention regions
Relation inference
Spatial anchors

ABSTRACT

Ensuring proper use of personal protective equipment (PPE) is essential for improving workplace safety management. The authors present an extensible pose-guided anchoring framework aimed at multi-class PPE compliance detection. The overall approach harnesses a pose estimator to detect worker body parts as spatial anchors and guide the localization of part attention regions using body-knowledge-based rules considering workers' orientations and object scales. Specifically, "part attention regions" are local image patches expecting PPEs based on their inherent relationships with body parts, e.g., (head, hardhat) and (upper-body, vest). Finally, the shallow CNN-based classifiers can reliably recognize both PPE and non-PPE classes within their corresponding part attention regions. Quantitative evaluations tested on the developed construction personal protective equipment dataset (CPPE) show an overall 0.97 and 0.95 F1-score for hardhat and safety vest detection, respectively. Comparative studies with existing methods also demonstrate the higher detection accuracy and advantageous extensibility of the proposed strategy.

1. Introduction

Detecting proper use of personal protective equipment is crucial for promoting safety management in construction workplaces. Construction sites continue to be among the most accident-prone and potentially hazardous workplaces [1]. Excessive risks (e.g., working at height, collapse, and manual handling) on the job site frequently expose workers to injuries and even fatalities. To prevent accidents, personal protective equipment (PPE) aims at protecting the wearer's body against job-related hazards. However, several factors, including low awareness, discomfort, fatigue, and carelessness, contribute to low compliance with PPE use and incorrect handling among workers [2]. Computer vision-based methods have shown potential for automated PPE compliance detection in past practices, as they permit non-invasive and low-cost perception on construction sites. Computer vision-based methods typically detect all workers and PPE components first and then verify if a worker uses the PPE based on spatial relationships among the workers and the involved PPE instances [3].

Although deep neural networks have led to significant progress in object detection, detecting individual workers and PPE items at construction sites remains a challenge due to the complex backgrounds around workers. Most object detection methods tend to scan the whole image to localize and classify multi-class objects, which may result in false detections on cluttered construction backgrounds. Another challenge in object detection for PPE and workers is the significant variations in object scales resulting from dynamic camera perspectives [4]. The shape and size of objects can change noticeably in videos when captured by cameras over varying distances. Furthermore, most existing methods for inspecting PPE compliance focus only on detecting hardhats. When verifying multi-class PPE objects, an accurate and robust object detection model requires collecting a large-scale domain-specific object dataset covering various scenarios for model training, which can be costly, tedious, and time-consuming.

Even after detecting individual objects, whether a worker lacks PPE remains to be verified. Many state-of-the-art approaches pair individual workers with their PPE by checking if the detected PPE is present in or around a worker's detection region [5,6]. However, these methods often fail to identify cases of incorrect PPE handling. For example, an employee may just hold the hardhat instead of wearing it on the head, as shown in Fig. 1a. Significant variations in workers' postures and orientations make it hard to enumerate all possible spatial relationships between the workers and PPE proposals. Additionally, these methods could hardly handle crowded workspaces where workers could occlude each other partially; bounding box representations (i.e., axis-aligned

* Corresponding author.
E-mail addresses: ruoxinx@andrew.cmu.edu (R. Xiong), ptang@andrew.cmu.edu (P. Tang).

https://doi.org/10.1016/j.autcon.2021.103828
Received 23 September 2020; Received in revised form 17 May 2021; Accepted 13 July 2021
Available online 22 July 2021

rectangles tightly bounding the object) used by these methods make it hard to isolate individual workers in crowded scenarios. The overlaps between bounding boxes could lead to matching confusion for individual workers and PPE items (see Fig. 1b). Furthermore, the computational complexity of spatial verification expands significantly with the number of PPE elements and workers [3]. For example, for two PPE instances and two workers, four ($2 \times 2 = 4$) possible combinations need spatial verification. However, for two PPE items and five workers, there are ten ($2 \times 5 = 10$) pairs of individual instances. The number of such combinations will grow exponentially and make the spatial verification of all of the instances unmanageable.

Recently, pose estimation has gained increasing attention in the computer vision community. The mission is to identify, localize, and track anatomical keypoints (also known as joints) for individuals in images or videos. Current pose estimation algorithms based on deep learning have achieved impressive results in unconstrained environments, demonstrating the potential for worker detection in complex construction environments. Contrasted with bounding-box based approaches for worker detection, human skeletons can provide more fine-grained information (e.g., location and visibility) about a person, especially in the occluded conditions (see Fig. 1b).

Motivated by the success of pose estimation models, the authors propose an extensible framework that leverages skeleton-based human pose information to improve multi-class PPE detection. First, the authors use a pose estimator rather than object detection methods to detect and represent individual workers in the form of human skeletons, which help isolate each worker from a crowded workspace where severe occlusions exist between workers. Empirical observations indicate that workers' poses also provide joint-level anchors for guiding the localization of different PPE items, e.g., (head, hardhat) and (upper-body, vest). This paper then defines human-body-part attention regions (the authors will use the term – "part attention regions" for the rest of the paper) that are informative image areas spatially correlated with PPE items. This attention-guided strategy can produce more accurate PPE detection results for two reasons: (1) the computational resource can be guided to concentrate on an informative local region, and (2) the local patch appearance can be shared between workers to benefit various backgrounds. To navigate through these image patches, the authors develop body knowledge-based rules using detected 2D keypoints to configure the location and size of the objects' bounding boxes under various workers' orientations. Finally, this study trains two shallow CNN-based classifiers to recognize hardhats or vests within cropped part attention regions. The efficient inference of non-PPE use workers is to identify those areas where the expected PPE is missing without evaluating the complex spatial relationships of the instances involved. To evaluate the performance of the proposed method, the authors introduce a new Construction Personal Protective Equipment (CPPE) Dataset and publicly release all data and annotations to encourage future research in the area. Extensive experiments verify this new method's potential by simultaneously checking for safety violations of non-hardhat use and non-vest use within the paper's scope.

In the remainder of this paper, Section 2 reviews the literature on the necessity for PPE use monitoring and computer vision-based PPE detection, as well as recent progress in pose estimation. The authors then describe the details of the proposed method in Section 3. Section 4 introduces the constructed CPPE dataset and describes the implementation details. Section 5 then evaluates the individual components of the developed method and compares this work against state-of-the-art approaches. Section 6 discusses the research limitations, followed by a summary of the research findings and future studies in Section 7.

2. Literature review

In this section, the authors first discuss the necessity of inspecting multiple PPE items in the workplace. Next, the authors review the main techniques for PPE detection at construction sites. Finally, since the authors recommend a pose-guided anchoring framework for PPE detection, a review of emerging human pose estimation methods is provided.

2.1. Importance of personal protective equipment (PPE) in construction

PPE acts as a fundamental barrier between workers and hazardous conditions in workplaces. Depending on the body-protected areas, standard PPE classifications and examples include head protection, eye and face protection, hand protection, body protection, foot protection, and hearing protection [7]. For instance, workers wearing hardhats can mitigate the impact of falling objects and avoid injuries from accidental bumps to stationary objects. Gloves are essential for shielding hands when handling rough or sharp materials. Likewise, the use of reflective safety vests could increase workers' visibility in workspaces, lowering the likelihood of struck-by accidents, especially in low-light or dark conditions.

Despite the high prevalence of hazardous working conditions, compliance with PPE use in workspaces remains low. The Occupational Safety and Health Administration (OSHA) stated that the lack of or improper use of PPE was one of the most violated OSHA standards during the 2019 fiscal year [8]. Statistics from the Bureau of Labor Statistics (BLS) revealed that nearly 84% of the workers sustaining head injuries from non-hardhat use, only 1% of almost 770 workers experiencing facial injuries were correctly wearing face protection, and the utilization rate of safety shoes was 23% among those workers who suffered foot injuries [9]. Companies and employers may also face significant fines of up to $12,934 per violation for PPE non-compliance [10]. Therefore, detecting non-compliance with requirements for using multi-class PPE is necessary for the workplace.

2.2. Computer vision-based PPE detection

There are two main techniques for verifying PPE compliance at construction sites: vision-based and sensor-based [5]. Wearable sensor-based methods focus on applying external location sensors and then analyzing the recorded signals to monitor compliance with the PPE use policy. Kelm et al. (2013) [11] introduced a mobile radio-frequency identification (RFID) device to determine if the workers' PPE use conformed to the corresponding safety regulations. Similarly, Li et al. (2017) [12] proposed a non-hardhat wear inspection system by attaching silicone pressure sensors to the hardhats' sweatbands. Kim et al. (2018) [13] used a three-axis accelerometer sensor to detect the proper use of safety helmets. Despite their potential to provide prompt

(a) hardhat use vs. non-hardhat use (b) bounding boxes vs. human poses

Fig. 1. Illustration of challenges in PPE detection.

alarms, sensor-based methods inevitably cause discomfort to the wearers over long working hours. The use of many wearable sensors can also lead to additional investment. Therefore, this paper focuses primarily on vision-based approaches for verifying PPE compliance in the workplace.

Conventionally, provided with a frame from the surveillance camera on construction sites, the vision-based techniques perform PPE compliance detection through two stages: object detection and relationship verification [3]. The first step is to detect workers and PPE items in the images. Previous works relied mainly on handcrafted features (e.g., shape, motion, color, and edge) to detect these objects. Wu et al. (2018) [14] utilized the histogram of oriented gradients (HOG) descriptor and support vector machine (SVM) for worker detection and then applied a color-based hybrid descriptor for hardhat identification. Mneymneh et al. (2019) [15] identified moving workers using background subtraction and detected the hardhat in the human head regions with the color-based classification algorithm.

Recently, deep learning techniques, such as Fast/Faster R-CNN [16,17], You Only Look Once (YOLO) [18], and Single Shot Detection (SSD) [19], have emerged as powerful methods for their exceptional machine learning abilities from large-scale labeled datasets. Fang et al. (2018a) [20] developed a Faster R-CNN method to detect non-hardhat use workers. Similarly, Fang et al. (2018b) [21] also utilized the Faster-R-CNN model to identify workers and their harnesses. Wu et al. (2019) [22] applied an SSD-based algorithm to identify workers with hardhats. Nath et al. (2020) [3] built on the YOLO architecture to verify non-compliance with hardhats and safety vests. However, these methods generally regard PPE detection as a specific target of object detection. Due to the cluttered backgrounds, the large variability of object scale, and common occlusions at construction sites, detecting multi-class PPE items will demand thousands or tens of thousands of domain-specific data samples for training these "data-hungry" methods.

When performing the relationship verification task, previous studies often relied on defining geometric and spatial rules to assess the contextual relationships of the detected instances of workers and PPE. For example, Park et al. (2015) [5] matched human bodies' windows and hardhats with predefined spatial rules. Nath et al. (2020) [3] verified if a worker was using a hardhat or vest by checking the Intersection over Union (IoU) of the bounding boxes that surround hardhats/vests and workers. Tang et al. (2020) [23] further designed a new human-object interaction (HOI) recognition method to check PPE compliance by detecting potential worker-PPE box pairs. Some researchers also integrate geometric rules to verify the proper use of PPE. Chen et al. (2020) [24] used the Euclidean distance between bounding boxes of detected hardhats and the neck to determine whether everyone uses the hardhat. However, the relationship verification's computational complexity expands with the number of PPE instances and workers because possible combinations of PPE instances and workers increase exponentially in response to those numbers [3]. Furthermore, the spatial relationships between workers and PPE may change with the workers' poses and orientations. That fact makes it difficult to define all possible verification rules.

Facilitated by recent achievements in face detection, a few researchers have attempted to utilize face regions to aid in non-hardhat use detection. For example, Du et al. (2011) [25] identified non-hardhat use workers by first detecting faces using Haar-like face features and then checking the presence of a hardhat based on color features around face regions. Shrestha et al. (2015) [26] also implemented an automatic non-hardhat use detection method by identifying the workers' faces and then using edge detection to localize hardhats near upper-head regions. Shen et al. (2020) [4] developed a face bounding-box regression algorithm to determine the candidate regions of safety helmets. However, these face-region-based methods fail to detect cases where the workers have their backs facing the camera. Additionally, most of the existing techniques are exclusively for detecting safety issues associated with hardhat use. Many other standard PPE components (e.g., safety vests or gloves) can hardly be applied to such face-based schemes.

2.3. Human pose estimation

"Human" is a special class in the computer vision community. Pose estimation explicitly represents "human" with human skeletons, which is a crucial step towards many domain applications, such as activity recognition [27], intelligent driver assistance systems [28], sign language understanding [29], and medical healthcare [30]. Classical pose estimation algorithms, such as pictorial structures [31] and deformable part models [32], have shown low detection accuracy in unconstrained environments.

With the introduction of the DeepPose network by Toshev and Szegedy (2014) [33], deep learning-based models have significantly reshaped human pose estimation techniques. The DeepPose network formulated the pose estimation as a joint regression problem using CNNs, which has yielded drastic improvements over standard benchmarks. Later, Wei et al. (2016) [34] designed a Convolutional Pose Machine network with iterative convolutional and pooling layers to output a set of heatmaps (also known as confidence maps) for keypoint prediction. Instead of regressing to XY locations, heatmaps model the joint distributions as Gaussian peaks. Newell et al. (2016) [35] proposed a Stacked Hourglass network using successive convolutional layers and residual modules. The test results on the FLIC dataset [36] reveal that the Stacked Hourglass network has obtained 99% Percentage of Correct Keypoints (PCK) accuracy on elbow joints and 97% on wrist joints. Cao et al. (2017) [37] developed a real-time OpenPose network by modifying Convolutional Pose Machines with the Part Affinity Fields (PAFs), which encode both location and orientation information of the limbs to aid in pair matching. This method has attained state-of-the-art accuracy results on the MS COCO Keypoints Challenge with a detection speed of 22 frames per second (FPS) on a single Nvidia GTX 1080 Ti machine [37].

In the construction domain, human pose estimation has gained increasing attention in various occupational tasks, such as worker behavior analysis [38,39], ergonomic analysis [40], and productivity assessments [41]. Liu et al. (2017) [42] applied CNNs to estimate worker poses in sequential images within unconstrained and cluttered environments. The experimental results achieved 91.7% PCKh@0.5 of all keypoints localization in the steel beam cutting task. Yan et al. (2017) [43] developed an ergonomic posture recognition technique for construction hazard prevention in 2D skeleton motion. The test results have demonstrated the feasibility of estimating worker poses with 2D ordinary cameras in the workplace. Given the success of pose estimation algorithms under real-world scenarios, this paper examines how to leverage skeleton-based human pose estimation techniques to enhance PPE detection accuracy with improved computational efficiency and reduced needs for training large-scale image samples.

3. Methodology

This section details the pose-guided framework for multi-class PPE detection in workspaces. In particular, the authors highlight the technical differences between the existing PPE detection strategies and the developed method in this study. To validate that the framework is extensible for detecting multiple PPE classes, the authors evaluate safety violations of hardhats and safety vests within the scope of this work.

3.1. Overview of the proposed framework

Fig. 2 illustrates the overall framework of the proposed method. Three parts collectively address the challenges related to efficient and effective PPE detection in workspaces: (1) worker pose estimation, (2) part attention localization, and (3) binary classification for PPE and non-PPE use. The authors first use a pose estimator to detect individual workers with occlusions. Part attention localization module utilizes

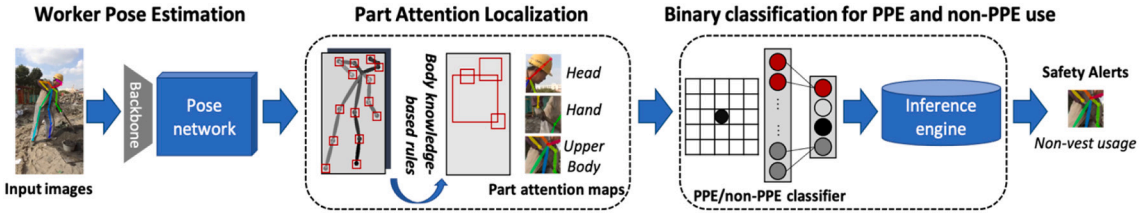


Fig. 2. Overview of the pose guided anchoring framework for multi-class PPE detection.

known 2D keypoints to locate informative image patches anticipating hardhats and vests. To guide through these part attention regions, the authors define body knowledge-based rules to configure the location and size of the cropping boxes. Finally, this study develops two CNN-based classifiers to determine whether these areas contain hardhats or vests. The algorithm then infers non-PPE use cases by identifying those anticipated PPE missing from their part attention regions.

The authors also summarize two widely adopted strategies from the existing literature and explain their technical differences with the proposed method, as are shown in Fig. 3. In scheme-1, the object detectors first identify all workers and PPE instances (in the form of a bounding box) in the images, then the spatial relations of workers and PPE are verified by evaluating the Intersection over Union (IoU) of identified bounding boxes. If these two bounding boxes (PPE box and worker box) overlap an intersecting area larger than a given threshold, then the worker is classified as correct PPE use [3,5,6]. To further reduce the search spaces, recent methods (scheme-2) will first localize workers' regions and then identify different PPE types within or around the bounding boxes of detected workers [3,14,15]. In contrast, the proposed method (scheme-3) uses a pose estimator to detect and represent individual workers in the form of skeletons. Instead of simply considering PPE detection as a specific application of object detection, the authors transform the process of PPE detection into a binary classification problem. To achieve this goal, the authors integrate the spatial anchors of worker poses to predetermine the candidate regions for PPE, e.g., (head attention region, hardhat) and (upper-body attention region, vest). This optimized strategy can effectively reject distracting backgrounds while improving PPE recognition accuracy within limited training samples.

3.2. Worker pose estimation

Instead of detecting workers with bounding box representations, the authors applied the pose estimation method for fine-grained detection and representation of workers' body parts. The pose estimation algorithm will identify all keypoints of the workers first and assemble these keypoints that belong to the same person in the image, which can provide joint-level information (e.g., location and visibility) about a person, especially in crowd scenes.

This study applies the OpenPose model developed by Cao et al. [37] for worker pose estimation for the following reasons: 1) well-established implementation for multi-person 2D pose detection with reliable results; and 2) real-time performance, which can attain near-real-time estimation in real-world conditions regardless of the number of people in any given image. To further speed up the detection process, the optimized method adopts the MobileNet network [44] rather than the original VGG-19 [45] as the feature extractor. The lightweight network uses depth-wise separable convolution filters that separate depth and spatial

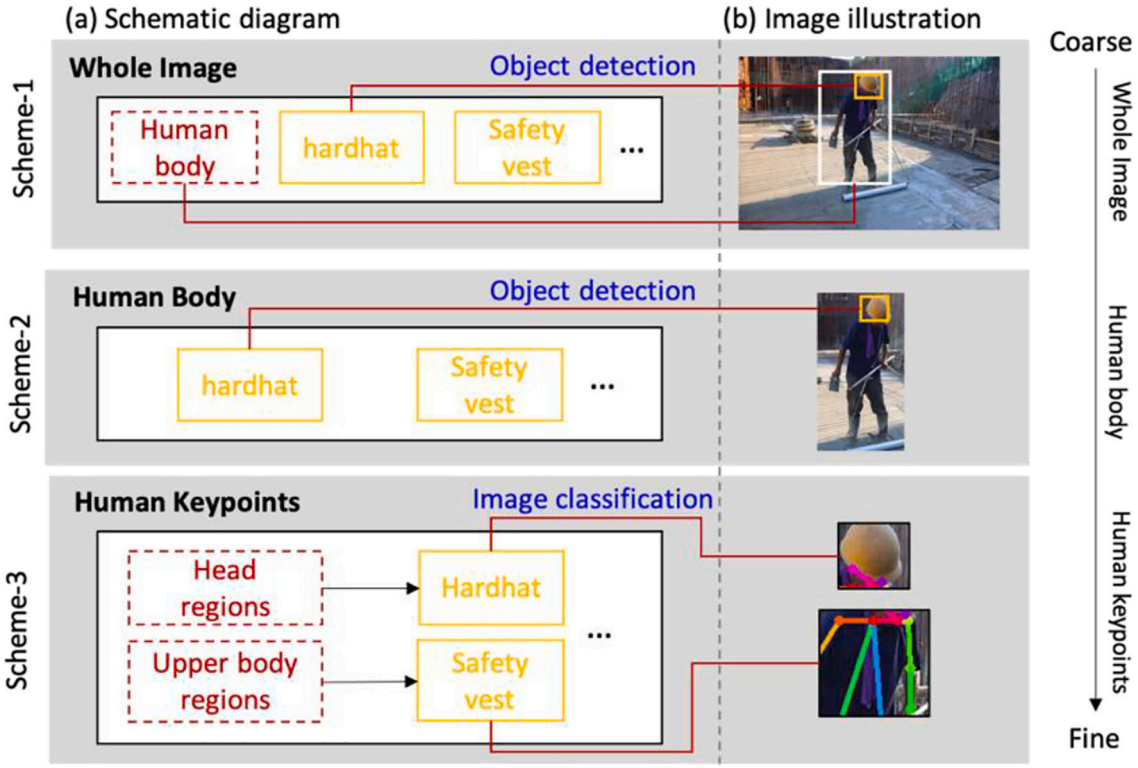


Fig. 3. Comparison of PPE detection strategies.

dimensions to improve computational efficiency.

Fig. 4 illustrates the pipeline of the lightweight OpenPose model. The network uses an iterative two-branch architecture, which simultaneously refines the pose estimation results over n consecutive stages. One branch generates a set of Part Affinity Fields (PAFs) for the pair-wise association, which connects all detected body parts to form full-body skeletons. PAFs link two associated keypoints of limbs and represent their locations and orientations with a list of 2D vector fields. The other branch produces coarse-to-fine heatmaps for part detection. Instead of an end-to-end coordinate regression, each heatmap is a 2D representation of the probability that a keypoint occurs at each pixel location, where a single peak implies its most likely location in the image. After that, the greedy inference can parse each heatmap and PAF map to assemble the candidate connections into full-body poses for multiple workers.

For model training, the authors compute the ground-truth heatmap $\mathbf{H}$ based on annotated keypoints. Let $\mathbf{x}_{i,j}$ be the ground-truth location of the i-th keypoint for the j-th worker in the image. Placing a 2D Gaussian distribution centered at $\mathbf{x}_{i,j}$, the probability value at pixel location $\mathbf{x}$ in individual heatmaps $\mathbf{H}_{i,j}$ is defined as:

$$\mathrm{H}_{i,j}(\mathrm{x}) = exp\left(- \left\|\mathrm{x}–\mathrm{x}_{i,j}\right\|_2^2 / \sigma^2\right) \quad (1)$$

Mathematically, the authors model individual workers with a list of pose skeletons. The set $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J)$ denotes all the poses of individual workers, where J is the total number of workers in the frame. For the j-th worker, $\mathbf{p}_{i,j} = (x_{i,j}, y_{i,j}, v_{i,j})$ denotes the predicted 2D coordinates of the i-th keypoint and its visibility $v_{i,j}$. For the human skeleton model, each pose has a total of $I = 18$ keypoints, including 1) nose, 2) neck, 3) right shoulder, 4) right elbow, 5) right wrist, 6) left shoulder, 7) left elbow, 8) left wrist, 9) right hip, 10) right knee, 11) right ankle, 12) left hip, 13) left knee, 14) left ankle, 15) right eye, 16) left eye, 17) right ear, and 18) left ear.

The detected keypoints can serve as spatial anchors for guiding attention to anticipated body part regions (i.e., part attention regions) depending on the types of PPE items, as illustrated in Table 1. For example, workers wear hardhats on the head to mitigate head impacts so that the keypoints such as the ears and nose can help navigate to the head regions. Similarly, shoulders and hips can predetermine potential regions for detecting safety vests, which are typically present around upper-body areas. Safety glasses protect the eye areas; ankles can initially localize potential areas for recognizing safety-toed footwear; wrists can guide gloves' detection; and ears are spatially relevant to earmuffs. For the scope of this research, the authors focus on detecting the proper use of hardhats and safety vests.

3.3. Part attention localization

The next step is to integrate the detected 2D keypoints as spatial anchors to infer and localize part attention regions, thus effectively eliminating the distracting backgrounds and guiding computational resources within informative local regions based on the PPE types.

Fig. 5 shows an overview of the part attention localization module. In this work, the authors examine two types of part attention regions: head attention regions (Region Type I) and upper body attention regions (Region Type II) to recognize expected hardhats and safety vests. Mathematically, this study formulates part attention regions as a set of bounding boxes $\mathbf{R} = \{\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_K\}$ for the candidate PPE items, where K is the total number of part attention regions in the image. To localize these part attention regions, the authors define body knowledge-based rules to configure the location and size of bounding boxes concerning diverse worker poses and orientations. Specifically, the following paragraphs detail two rules used to assign dynamic head and upper body attention regions for PPE recognition.

Table 1
Spatial anchors for localizing PPE items.

PPE items	Body protection	Spatial anchors
hardhat	head	ears; nose
safety vest	body	shoulders; hips
safety goggles	eyes	eyes
safety-toe footwear	feet	ankles
gloves	hands	wrists
earmuffs	hearing	ears

(1) Head attention regions for hardhat recognition.

Head attention regions (Region Type I) are candidate areas in the image for hardhat recognition. For each Region Type I, the authors select the *nose* and *ears* as the reference points to determine the location and size of Region Type I for a given worker instance. The redundant joints guarantee cropping performance when some keypoints are invisible in the image. Considering the workers' relative orientation to the camera, the visibility of *nose* and *ears* consists of five situations: a) ears are visible, but the nose is invisible, b) one ear and nose are visible, c) ears and nose are visible, d) one ear is visible while the nose is invisible, e) ears and nose are invisible. Fig. 6 shows these cases of visible joints for defining head attention regions.

Let x_{le}, y_{le}, x_{re}, y_{re}, x_n, y_n, x_{neck}, y_{neck} denote the coordinates of the *left ear*, *right ear*, *nose*, and *neck*, respectively. v_{le}, v_{re}, $v_n \in \{0,1\}$ represent the visibility of the *left ear*, *right ear*, and *nose* in the same way.

Case 1-A - ears are visible. In this case, workers present with the back view. This study utilizes oriented bounding boxes to represent Region Type I in the format of $(x_1, y_1, l_1, \theta_1)$, where x_1, y_1 denote the midpoint coordinates on the bottom edge, l_1 is the side length, and θ_1 is rotation angle. In this study, the authors used oriented region proposals rather than horizontal bounding boxes to localize part attention regions. These rotated image patches with additional angle parameters are generated adaptively according to the workers' orientations, which helps describe their locations and contents more accurately than axis-aligned boxes. The rules determine the expected regions of hardhats as follows:

$$x_1 = (x_{le} + x_{re})/2$$

$$y_1 = (y_{le} + y_{re})/2$$

$$l_1 = r_a\left[(x_{le}–x_{re})^2 + (y_{le}–y_{re})^2\right]^{1/2}$$

$$\theta_1 = tan^{-1}\left[(y_{le}–y_{re})/(x_{le}–x_{re})\right] \quad (2)$$

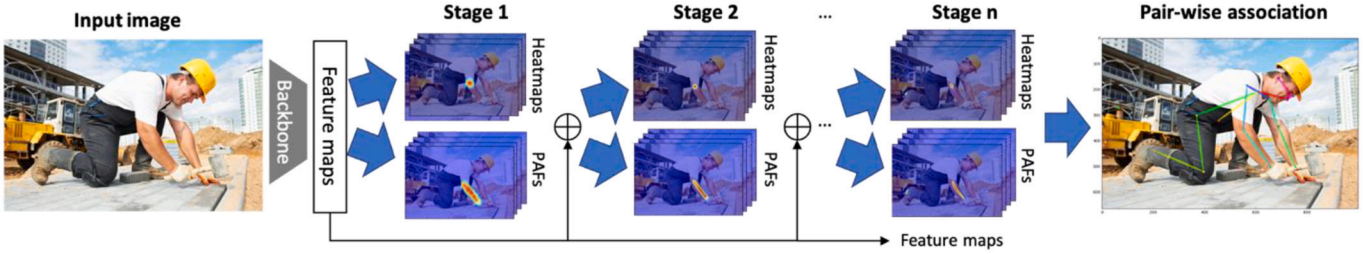


Fig. 4. Architecture of the OpenPose network.

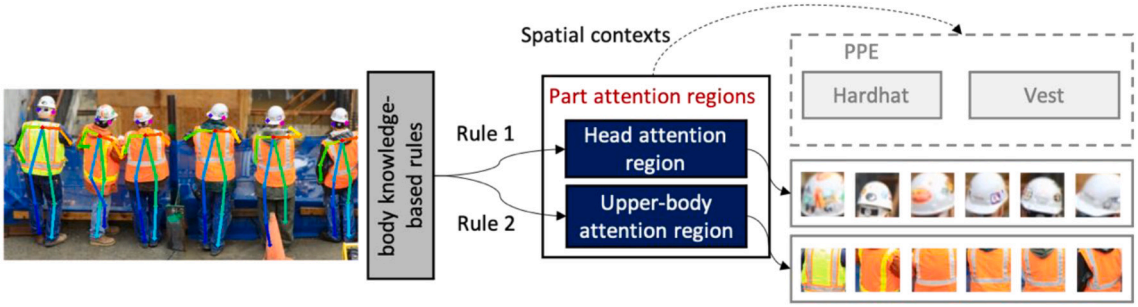


Fig. 5. Overview of the part attention localization module.

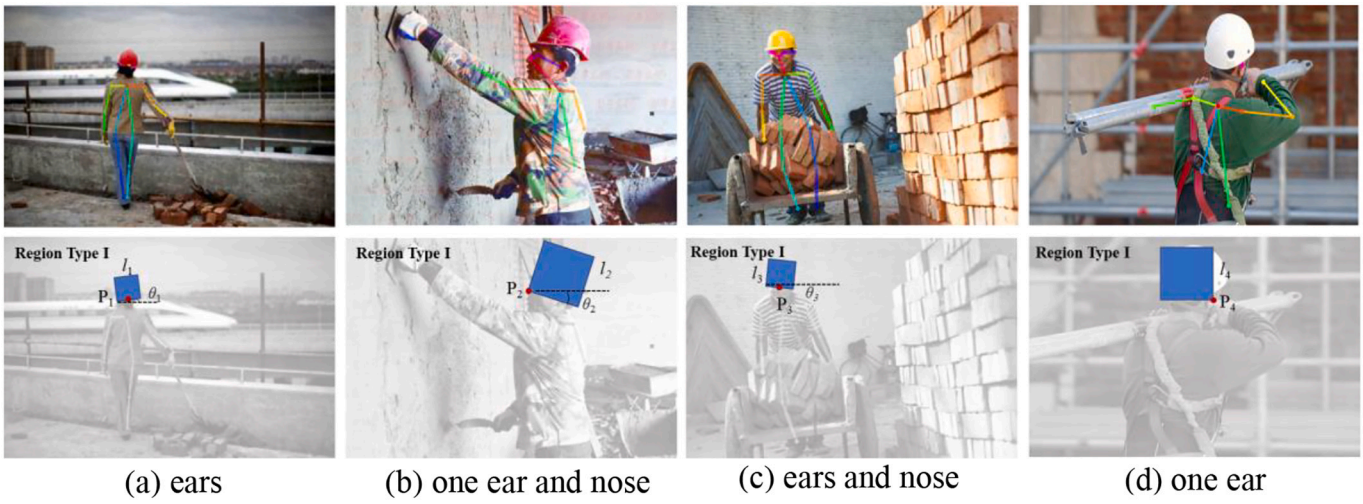


Fig. 6. Illustration of the body knowledge-based rules for configuring head attention regions.

where r_a is a hyperparameter to ensure that the local window scales with the workers' size.

Case 1-B - one ear and nose are visible. In this case, workers appear in a side view. Similarly, the rules segment the expected regions as follows:

$$x_2 = \mathrm{x_n}$$

$$y_2 = \mathrm{y_n}$$

$$l_2 = r_b\left[(x_n - \mathrm{x_{neck}})^2 + (y_n - \mathrm{y_{neck}})^2\right]^{1/2}$$

$$\theta_2 = tan^{-1}\left[(\mathrm{y_{le}v_{le}} + \mathrm{y_{re}v_{re}} - y_n)/(\mathrm{x_{le}v_{le}} + \mathrm{x_{re}v_{re}} - x_n)\right] \quad (3)$$

where x_2, y_2 denote the corner coordinates on the bottom edge, l_2 is the side length, and θ_2 is rotation angle $\mathrm{r_b}$ is a hyperparameter to regulate the proper size of Region Type I.

Case 1-C - ears and nose are visible. In this case, the worker poses are present in a front view. The rule for calculating the oriented box is the same as **Case** 1-A.

Case 1-D - one ear is visible while the nose is invisible. Since no reference point is visible in the regions, the rotation angle $\theta_4 = 0$ in this case. The rules for localizing the head areas are as follows:

$$\mathrm{x_4} = \mathrm{x_{le}v_{le}} + \mathrm{x_{re}v_{re}} \quad (4)$$

$$\mathrm{y_4} = \mathrm{y_{le}v_{le}} + \mathrm{y_{re}v_{re}}$$

$$l_4 = r_d\left[(\mathrm{x_{le}v_{le}} + \mathrm{x_{re}v_{re}} - \mathrm{x_{neck}})^2 + (\mathrm{y_{le}v_{le}} + \mathrm{y_{re}v_{re}} - \mathrm{y_{neck}})^2\right]^{1/2}$$

where x_4, y_4 denote the corner coordinates on the bottom edge, and l_2 is the side length. r_d is a hyperparameter to determine the proper size of Region Type I.

Case 1-E - ears and nose are invisible. Severe head occlusions can lead to head invisible cases, where none of the keypoints within the head regions are visible in the image.

(2) Upper-body attention regions for vest recognition.

Upper-body attention regions (Region Type II) are potential regions that expect the existence of safety vests. For Region Type II, the authors select *shoulders* and *hips* as the reference points to determine the location and size of Region Type II. Considering their orientation to the camera, the visibility of the *shoulders* and *hips* contains five cases: a) shoulders and hips are visible, b) shoulders are visible while hips are invisible, c) shoulders and one hip are visible, d) one shoulder is visible, and e) shoulders are invisible. Fig. 7 illustrates these possible cases of worker poses for defining upper-body attention regions.

Let x_{ls}, y_{ls}, x_{rs}, y_{rs}, x_{lh}, y_{lh}, x_{rh}, y_{rh} denote the coordinates of the *left shoulder*, *right shoulder*, *left hip*, and *right hip*, respectively. v_{ls}, v_{rs}, v_{lh}, and v_{rh} represent the visibility of the *left shoulder*, *right shoulder*, *left hip*, and *right hip*, respectively.

Case 2-A -s houlders and hips are visible. The authors represent the oriented bounding boxes in the format of (x_1, y_1, w_1, h_1, θ_1), where x_1, y_1, w_1, h_1, θ_1 denote the midpoint coordinates on the upper edge, width, height, and rotation angle of the bounding boxes, respectively. The rules for localizing the expected regions of safety vests are as follows:

$$x_1 = (x_{ls} + x_{rs})/2 \quad (5)$$

$$y_1 = (y_{ls} + y_{\mathrm{rs}})/2$$

$$h_1 = \left[(y_{ls} + y_{rs} - y_{lh} - y_{rh})^2/4 + (x_{ls} + x_{rs} - x_{lh} - x_{rh})^2/4\right]^{1/2}$$

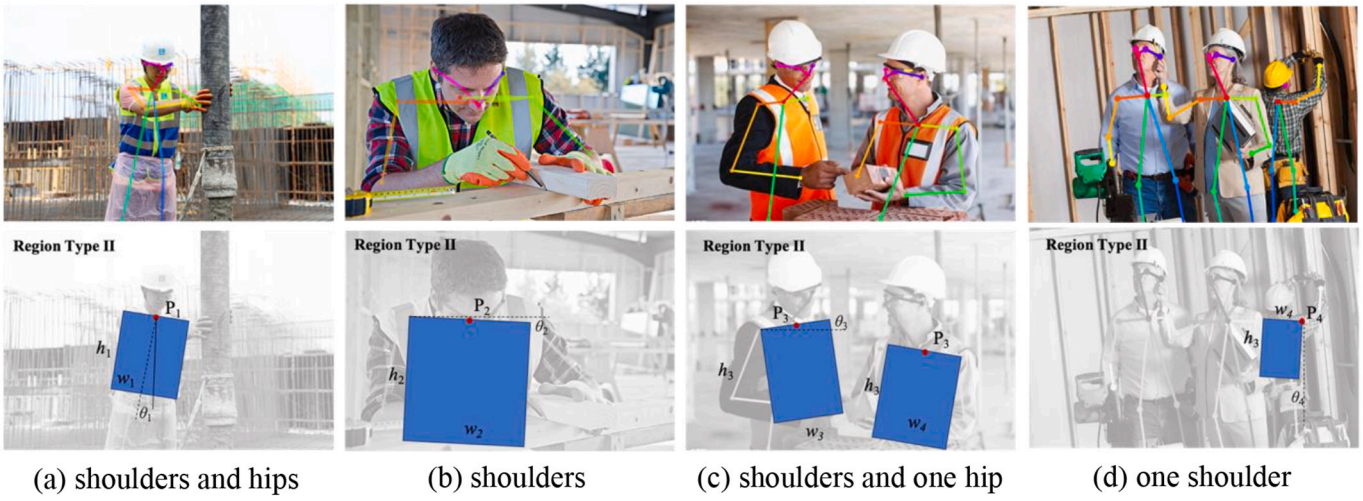


Fig. 7. Illustration of the body knowledge-based rules for configuring upper-body attention regions.

$$w_1 = s_a\, h_1$$

$$\theta_1 = tan^{-1}[(y_{ls} + y_{rs} - y_{lh} - y_{rh})/(x_{ls} + x_{rs} - x_{lh} - x_{rh})]$$

where s_a is the side length ratio of rotated boxes for Region Type II.

Case 2-B -s houlders are visible while hips are invisible. The rules for cropping the expected regions of safety vests are as follows:

$$x_2 = (x_{ls} + x_{rs})/2 \tag{6}$$

$$y_2 = (y_{ls} + y_{rs})/2.$$

$$w_2 = \left[(y_{rs} - y_{ls})^2 + (x_{rs} - x_{ls})^2\right]^{1/2}$$

$$h_2 = s_b\, w_2$$

$$\theta_2 = tan^{-1}[(y_{rs} - y_{ls})/(x_{rs} - x_{ls})]$$

where $x_1, y_1, w_1, h_1, \theta_1$ denote the midpoint coordinates on the upper edge, width, height, and rotation angle of the bounding boxes, respectively. s_b is the side length ratio of rotated boxes for Region Type II.

Case 2-C – shoulders and one hip are visible. In this case, the rule for calculating the oriented box is the same as **Case 2-B**.

Case 2-D – one shoulder is visible. The rules for determining upper-body attention regions are defined as follows:

$$x_4 = x_{ls}\, v_{ls} + x_{rs}\, v_{rs} \tag{7}$$

$$y_4 = y_{ls}\, v_{ls} + y_{rs}\, v_{rs}$$

$$h_4 = \left[(y_{ls}\, v_{ls} + y_{rs}\, v_{rs} - y_{lh}\, v_{lh} - y_{rh}\, v_{rh})^2 + (x_{ls}\, v_{ls} + x_{rs}\, v_{rs} - x_{lh}\, v_{lh} - x_{rh}\, v_{rh})^2\right]^{1/2}$$

$$w_4 = s_d\, h_4$$

$$\theta_4 = tan^{-1}[(y_{ls}\, v_{ls} + y_{rs}\, v_{rs} - y_{lh}\, v_{lh} - y_{rh}\, v_{rh})/(x_{ls}\, v_{ls} + x_{rs}\, v_{rs} - x_{lh}\, v_{lh} - x_{rh}\, v_{rh})]$$

where $x_4, y_4, w_4, h_4, \theta_4$ indicate the corner coordinates on the upper edge, width, height, and rotation angle of the bounding boxes, respectively. s_d is the side length ratio of rotated boxes for Region Type II.

Case 2-E - shoulders are invisible. Severe occlusion occurs in upper body regions, so the authors define these cases as upper-body invisible.

Fig. 8 shows several examples of localized head and upper body attention regions. By cropping these part attention regions, even if approximate, the proposed localization strategy could efficiently reduce the search spaces while distributing computational resources on small human-body areas that are candidates of PPE items.

Fig. 8. Part attention regions obtained for hardhat and vest recognition.

3.4. Binary classification for PPE and non-PPE use

Workers wear the PPE items to protect a specific body area properly. The inherent relationships between the PPE instances and local body parts, e.g., (head, hardhat) and (upper-body, vest), can be used to aid in inferring the fact whether the worker is using the PPE without a detailed analysis of the spatial relations of the objects involved. Therefore, the classification of the cases into PPE use and non-PPE use classes relies on detecting PPE instances in the corresponding body part attention region. For example, if a hardhat appears in the head attention region, the worker is regarded as correctly complying with the compliance requirements of hardhats. In other cases, if the expected hardhat is missing from the head attention regions, the corresponding worker will be labeled as the NH.

Each worker can produce two types of part attention regions: head attention regions and upper-body attention regions if their body parts are visible in the image. To recognize PPE instances within the part attention regions, the authors develop two classifiers: hardhat classifier $f_1(\mathbf{X}_1)$ and vest classifier $f_2(\mathbf{X}_2)$. Specifically, head attention regions comprise two classification results- hardhat use (WH) and non-hardhat use (NH), while the upper-body attention regions have two classes - vest use (WV) and non-vest use (NV). Based on the above analysis, the authors have implemented an inference engine to investigate the relationships between workers and PPE items and ultimately determine the categories of the focused regions as follows:

Input: $\mathbf{E}_{j,m}$ is the m-th ($m = 1, 2$) part attention region of the j-th worker.
Output: The PPE-use labels $\boldsymbol{o}_{j,k}$ of the j-th worker.
$\boldsymbol{o}_{j,k} = \emptyset$;
for each head attention region $\mathbf{E}_{j,1}$ of the j-th worker **do**:
  Apply the hardhat classifier $f_1(\mathbf{X}_1)$ to predict the hardhat label $o_{j,1}$ (i.e., WH or NH);
  Add the label $o_{i,1}$ to $\boldsymbol{o}_{j,k}$;
end
for each upper body attention region $\mathbf{E}_{j,2}$ of the j-th worker **do**:
  Apply the vest classifier $f_2(\mathbf{X}_2)$ to predict the vest label $o_{j,2}$ (i.e., WV or NV);
  Add the label $o_{j,2}$ to $\boldsymbol{o}_{j,k}$;
end
Return $\boldsymbol{o}_{j,k}$

For image classification, the prevalent deep CNN networks such as VGG [45], Inception [46], and ResNet [47] have attained impressive recognition results on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [48]. However, due to their network complexity with many parameters, very deep neural networks may have difficulty in optimizing the parameters and are prone to get overfitting during the training process. Given the relatively small size of cropped image patches (usually 32×32 or 64×64 in pixels) in this study, the authors adopted a shallow PPE/non-PPE classifier for hardhat and vest recognition based on the improved LeNet architecture [49].

Fig. 9 illustrates the architecture of the proposed classifier. This CNN classifier consists of six layers, including iterative convolutional and subsampling layers, along with fully connected layers. To illustrate, the authors denote convolutional layers (C layers), subsampling layers (S layers), and fully connected layers (F layers) as C_x, S_x, and F_x, respectively, where x refers to the layer index.

The architecture takes the cropped image patches as input. The first convolutional layer C1 produces twenty feature maps from 5 × 5 filters. The feature maps in the C1 layer use different sets of weight parameters and biases, thus extracting multiple features from each location. The layer S2 with the filter size 2 × 2 and a stride of 2 can reduce the feature maps' height and width by half while the depth remains unchanged. Similar to C1, layer C3 is a convolutional layer with 5 × 5 kernels and fifty filters, resulting in fifty feature maps. The output of Layer C3 passes through the next subsampling Layer S4, which produces 16 feature maps. The fifth layer F5 is a fully connected layer containing 500 output units. Finally, the output layer F6 with Softmax activation assigns each input image into one of two classes (WH or NH; WV or NV).

To determine the normalized size of input images, the authors analyze the scale distributions of part attention regions in the training subset, as shown in Fig. 10. For head attention regions, the image patches in the scale range of (14, 40) account for more than 50% of the samples. Therefore, the resized window size of head attention regions is 32 × 32 pixels. Similarly, the cropped upper-body attention regions in the scale range of (29, 77) constitute more than half of the instances. The input resolution of the vest classifier is 64 × 64 pixels.

4. Dataset and implementation details

This section introduces the developed Construction Personal Protective Equipment (CPPE) dataset and describes the implementation details and experimental settings of training and validation steps.

4.1. Dataset statistics

Publicly available image datasets on evaluating proper use of personal protective equipment (PPE) involve Pictor-v3 dataset [3], GDUT-Hardhat Wearing Detection (GDUT-HWD) dataset [22], and Safety helmet wearing detect dataset (SHWD) [50]. These datasets have contributed to encouraging progress in ensuring site safety. However, GDUT-HWD and SHWD are established only for hardhat wearing detection. The public portion of the Pictor-v3 dataset contains few training samples that capture safety vest use scenarios. For this reason, this study introduces a new Construction Personal Protective Equipment (CPPE) dataset by collecting high-quality data from public datasets and web-mined images. The constructed CPPE dataset consists of 932 images, including 2747 instances of hardhats, 1339 instances of safety vests, and 3428 workers, while covering various construction activities, illuminations, occlusions, and resolutions (see Fig. 11).

The CPPE dataset contains 627 randomly selected training images and 305 testing images. Fig. 12 shows the number of hardhats, vests, hardhat use workers, non-hardhat use workers, vest use workers, and non-vest use workers in the training and testing subsets. The training subset consists of 2406 workers, 1890 instances of hardhats, 889 instances of vests, 1797 instances of hardhat use workers, 594 instances of non-hardhat use workers, 889 instances of vest use workers, and 1485 instances of non-vest use workers. The number of hardhats is larger than the number of hardhat use workers because some workers are not properly wearing the hardhats (see **Fig. 1**a). The testing subset consists of 1022 workers, 857 instances of hardhats, 450 instances of vests, 804

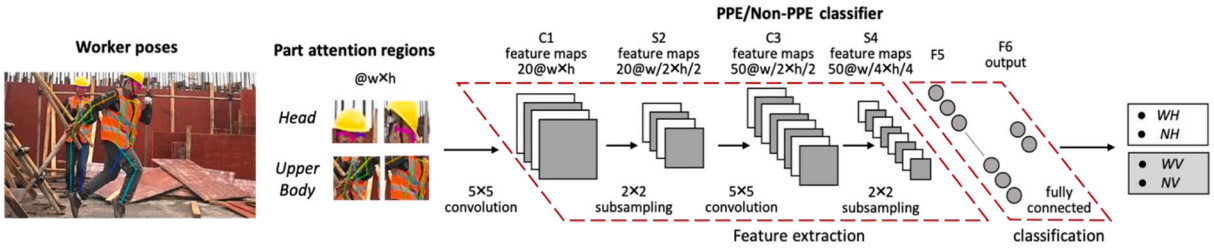
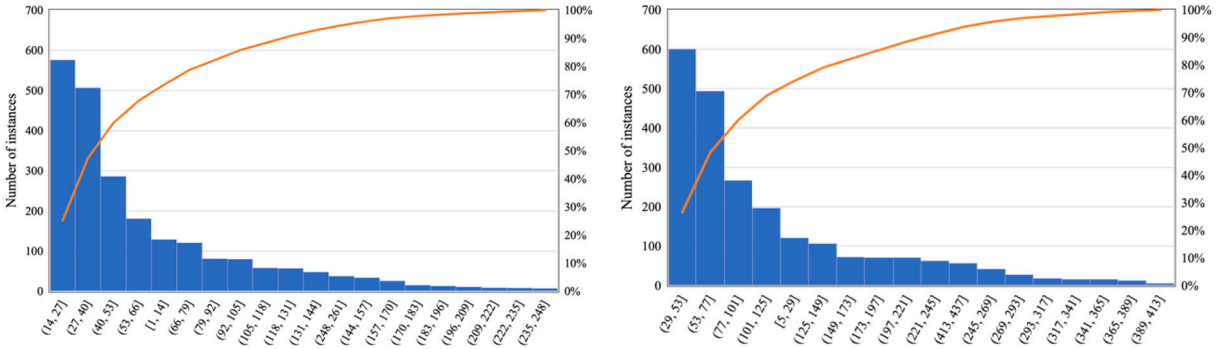


Fig. 9. The architecture of the PPE/non-PPE classifier network.



(a) Scale distributions of head attention regions (b) Scale distributions of upper-body attention regions

Fig. 10. Scale distributions of part attention regions.

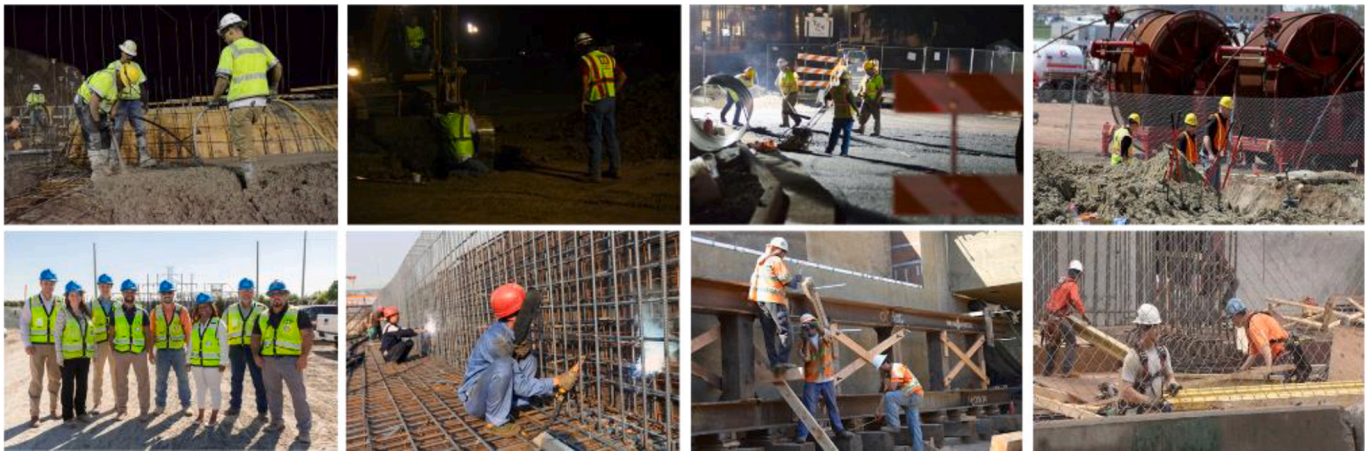


Fig. 11. Sample images from the CPPE dataset.

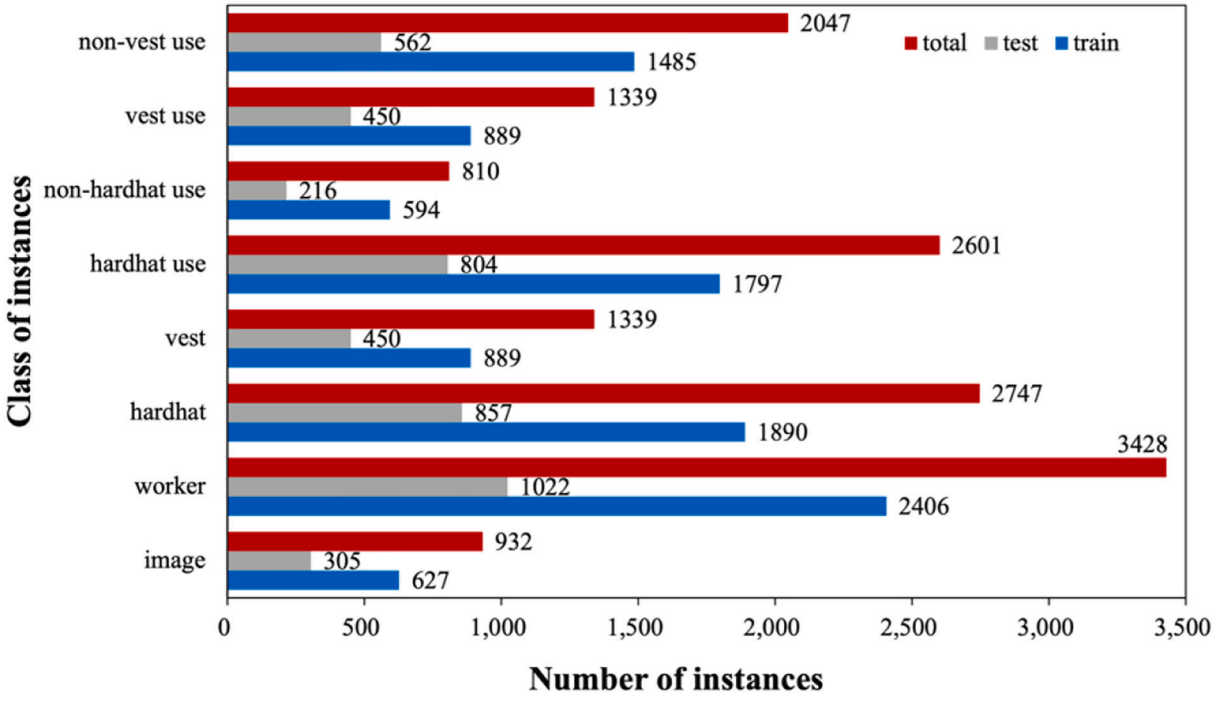


Fig. 12. Data distributions of the CPPE dataset.

instances of hardhat use workers, 216 instances of non-hardhat use workers, 450 instances of vest use workers, and 562 instances of non-vest use workers.

4.2. Implementation details

The proposed method follows a three-step framework. First, a pose estimator extracted joint-level information for individual workers. Second, the authors used the spatial anchors of worker poses to localize body part attention regions that anticipate PPE items. Finally, two developed image classifiers can recognize PPE instances within the cropped part attention regions.

4.2.1. Training of the lightweight OpenPose network

To speed up the inference process, the authors used the MobileNet network [44] rather than the original VGG-19 [45] as the backbone. The lightweight OpenPose network [37] pre-trained on the MS COCO keypoint train2017 split dataset [51] obtained the pose information of individual workers. Similar to the original work of OpenPose, the authors applied random cropping, rotation (± 45°), scaling (± 20%), and flipping (50%) for data augmentation. The input resolution is 368 × 368 pixels with a batch size of 10 images. The network used Adam optimizer (initial learning rate of 1e-4 and epsilon of 1e-9) for 440,000 training iterations on a server with two NVIDIA GeForce GTX 1080Ti GPUs.

4.2.2. Localization of the part attention regions

The authors localized two types of part attention regions: head attention regions (Region Type I) and upper body attention regions (Region Type II) with joint-level information from the deriving poses of workers. Considering the workers' relative orientation, the authors defined body knowledge-based rules to localize corresponding part attention regions in different poses. To test optimal parameters r and s for regulating the proper size of part attention regions, the authors randomly selected about 20% images from the training subset for extensive trials. Table 2 lists the scaling parameters of part attention regions used in the experiment.

4.2.3. Training of the PPE classifiers

After cropping the body part (head and upper body) regions from the whole image, the authors manually annotate image patches with two-class labels (i.e., WH or NH for head attention regions, and WV or NV for upper-body attention regions). To fine-tuning network parameters of hardhat classifier $f_1(\mathbf{X}_1)$, 20% of WH and NH image patches cropped from the training subset of the CPPE dataset are randomly selected as the validation subset. The localized head attention regions are 32 × 32 resized image patches. The batch size of all training models is 8, with the Adan optimizer and an initial learning rate of 1e-4 for 100 epochs. Similarly, the WV and NV image patches cropped from the training subset of the CPPE dataset are randomly split into 80% training samples and 20% validation samples for training the vest classifier $f_2(\mathbf{X}_2)$ with a batch size of 4. The cropped upper-body attention regions are informative image patches with a window size of 64 × 64. All classifiers, i.e., hardhat classifier $f_1(\mathbf{X}_1)$ and vest classifier $f_2(\mathbf{X}_2)$, are trained with Adam optimizer with an initial learning rate of 1e-4 for 100 epochs. The authors also decrease the learning rate by half every ten epochs. The data augmentation involved horizontal flipping, scaling (± 20%), rotation (± 30°), horizontal and vertical shifting (± 10%), and shearing (± 20%). The authors select the trained model with the highest accuracy in the validation subset for testing. In the experiment, the highest accuracy of hardhat classifier $f_1(\mathbf{X}_1)$ is at the 70th epoch, while the high accuracy of vest classifier $f_2(\mathbf{X}_2)$ is at the 46th epoch.

Table 2
Scaling parameters of part attentions regions.

Regions	Head attention regions			Upper-body attention regions		
Case	1-A ears	1-B one ear and nose	1-D one ear	2-A shoulders and hips	2-B shoulders	2-D one shoulder
Parameters	r_a	r_b	r_d	s_a	$s_b = 1/s_a$	$s_d = 1/s_a$
Values	1.2	1.5	1.0	0.6	1.7	1.7

5. Results and evaluations

This section evaluates the performance of individual elements of the proposed method. These elements support worker detection, part attention localization, and PPE recognition as needed by the framework. The authors also compare the proposed method with state-of-the-art approaches.

5.1. Performance evaluation of worker detection

The authors use precision, recall, and F1-score as evaluation metrics to evaluate worker detection performance. The precision and recall are defined as follows:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{8}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{9}$$

where true positive (TP) is the number of correctly detected workers, false positive (FP) is the number of detected workers that are actually non-workers, and false negative (FN) is the number of missing workers.

To measure the balanced performance of worker detection, the authors also use the F1 measure, which is the harmonic mean between precision and recall, as in Eq. (10):

$$F1 = 2 \times Precision \times Recall/(Precision + Recall) \tag{10}$$

Table 3 summarizes the precision and recall results for worker detection. The proposed method for worker detection achieved a 99.61% precision and a 98.04% recall in worker detection, meaning that 0.39% of the workers were incorrectly detected, and the algorithm missed 1.96% of the workers in the image. To evaluate the effect of scale variations of workers, the authors divide the CPPE dataset into three categories: small (0–96 pixels), medium (96–128 pixels), and larger (>128 pixels), based on worker heights (as given by the bounding box annotation). The test results (Table 3) show that the overall precision and recall of worker detection were above 95% for medium and large-scale cases. However, the proposed method showed a relatively low recall (71.97%) for worker detection under small-scale scenarios because extracting features from tiny persons is hard, and even human inspectors have difficulty in recognizing tiny instances from a long-range view.

The authors have analyzed the typical examples of false worker detections in the CPPE dataset – these images cause false positive or false negative cases, as shown in Fig. 13. In these examples, ambiguous objects at construction sites, such as humanoid shadows and equipment structures, may be incorrectly detected as human bodies in image frames. Introducing more negative instances helps the model discriminate between workers and other site objects, thus reducing false-positive cases. Moreover, as indicated in Table 3, the small scale of workers that often occur in practical workplaces also frequently leads to false-negative errors.

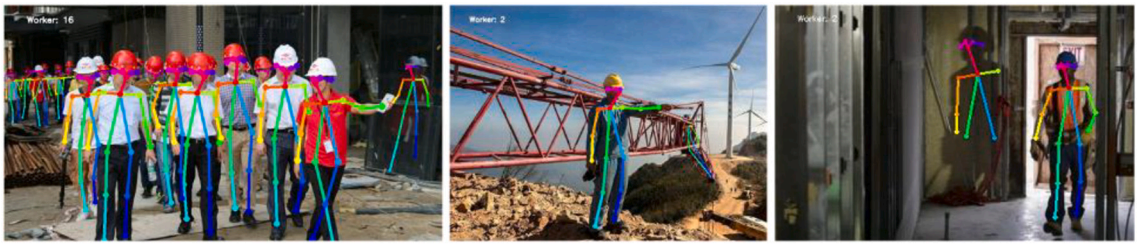
5.2. Performance evaluation of part attention localization

To assess the performance of part attention localization, the authors used the intersection-over-union (IoU) between the cropped part attention regions and ground-truth bounding boxes as the evaluation metrics. The calculation of IoU between oriented bounding boxes is similar to that between horizontal bounding boxes. The only difference is that the IoU calculation for oriented bounding boxes is performed

Table 3
Summary of the results for worker detection.

Task	Category	TP	FN	FP	Precision	Recall	F1-score
Worker detection	Small (0–96 pixels)	113	44	7	94.17	71.97	81.59
	Medium (96–128 pixels)	129	6	1	99.23	95.56	97.36
	Large (>128 pixels)	3119	17	5	99.84	99.46	99.65
	Total	3361	67	13	99.61	98.04	98.82

Note: precision, recall, and F1-score values are in percentage.



(a) FPs for worker detection

(b) FNs for worker detection

Fig. 13. Examples of false worker detections.

within polygons, as illustrated in Fig. 14.

The computation of the IoU between two oriented bounding boxes is as follows:

$$IoU = \frac{area(B_1 \cap B_2)}{area(B_1 \cup B_2)} \tag{11}$$

where B_1 and B_2 are two oriented bounding boxes.

The authors used the open-source tool roLabelImg [52] to label the oriented ground-truth boxes of part attention regions. In general, the higher the IoU of part attention regions is, the higher the localization accuracy is. The IoU will check whether the IoU between these two bounding boxes is higher than a defined threshold. For the task of part attention localization, *TP* is the number of correct localizations with an IoU ≥ 0.5. *FP* is the number of improper localizations with an IoU < 0.5, while *FN* is the number of ground truth regions not detected.

To further assess the impact of occlusion on part attention location, the authors classify the occlusion degree into different categories based on the number of visible anchoring keypoints in their corresponding part attention regions. Accordingly, the occlusion degree of head attention regions consists of five categories: “ears and nose,” “one ear and nose,” “ears,” “one ear,” and “head invisible.” The occlusion degree of upper-body attention regions contains four classes: “shoulders and hips”, “shoulders”, “one shoulder and one hip”, and “upper-body invisible”. In particular, the cases of “part invisible”, i.e., “head invisible” and “upper-body invisible”, will not generate any bounding boxes. *TP* is the number of correctly “part invisible” instances, *FP* is the number of “part invisible” instances that are actually “part visible”, and *FN* is the number of mis-detected “part invisible” instances.

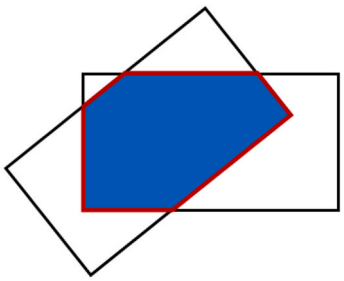


Fig. 14. Examples of IoU between oriented bounding boxes.

Table 4
Summary of the results for part attention localization.

Regions	Category	Number of anchoring keypoints	Precision	Recall	F1-score
Head	ears and nose	3	98.68	99.02	98.85
	one ear and nose	2	99.26	98.38	98.82
	ears	2	96.34	95.18	95.76
	one ear	1	87.88	76.32	81.69
	head invisible*	0	76.92	66.67	71.43
	Total		98.55	98.02	98.28
Upper body	shoulders and hips	4	99.68	98.81	99.24
	shoulders	2	91.89	94.01	92.94
	one shoulder and one hip	2	84.21	74.41	79.01
	upper-body invisible*	0	83.87	72.22	77.61
	Total		98.85	98.03	98.44

Note: precision, recall, and F1-score values are in percentage. * indicates that the performance evaluation metrics of “head invisible” and “upper-body visible” differ from other cases.

The test results (see Table 4) show that the overall precision and recall of hardhat and upper-body region localization are above 98%. However, upper-body region localization performance is higher than the performance of head region localization because the head regions are typically much smaller than the upper-body regions. The anchoring keypoints within the head attention regions have denser distributions than the anchoring keypoints of upper-body attention regions, which makes it challenging for the pose estimator to infer their locations accurately. The results also reveal that body occlusions can affect the performance of part attention localization. For example, with fewer anchoring keypoints, the precision and recall of head part region localization provide 87.88% and 76.32% results, and precision and recall of upper-body part region localization yield 84.21% and 74.41% performance.

The authors examined the typical cases where the proposed method fails in the CPPE dataset. Fig. 15 shows typical examples of false localization errors. Since the part attention regions are determined based on the detected worker poses, the localization errors and pose estimation failures are closely related. False part detection (Fig. 15a), which results from body occlusion, can lead to localization errors for head and upper body attention regions. High overlapping scenarios (Fig. 15b) can also lead to localization errors. In highly crowded scenes where workers are overlapping, the pose estimator may merge keypoints among different workers and partly miss detections. Fig. 15c shows false positives resulting from incorrect worker detections, while Fig. 15d refers to false-negative errors where the pose estimator fails to detect workers in the workplace.

5.3. Performance evaluation of PPE recognition

As for the PPE recognition, *TP* is the total number of the correctly classified cases where workers are using PPE, *FP* is the number of workers who are not using PPE properly is incorrectly identified as PPE use, *FN* is the number of workers who are wearing PPE is predicted as non-PPE use. In particular, the errors caused by incorrect part attention localization will not be calculated in the experiments since the objective of this section is to evaluate the PPE classifier's performance individually. The authors also compared the developed classifier with the state-of-the-art CNN classifiers [44,45,47], including VGG-16, VGG-19, MobileNet, ResNet-18, ResNet-34, ResNet-101, and ResNet-152. The runtime speed for each CNN classifier uses Frames Per Second (FPS) as the evaluation metrics by averaging the inference time on the testing subset.

The test results (see Table 5 and Table 6) show that VGG-19 and VGG-16 achieve the highest F1-score (0.98 and 0.97) on hardhat and vest recognition, respectively. The proposed method achieves a 97.13% precision, a 97.74% recall for hardhat recognition, and a 96.12% precision, a 94.61% recall for vest recognition. However, the shallow CNN classifier with far fewer parameters used by this study provides 179.3 FPS on the hardhat testing subset and 156.6 FPS on the vest testing subset, which is much faster than other methods. Furthermore, the CNN classifiers with different depths have shown similar classification performances. Two main reasons are: (1) the model inputs are typically low-resolution patches (usually 32×32 or 64×64 in pixels); and (2) the part attention regions retain the informative regions while eliminating distracting backgrounds. Consequently, the shallow CNN classifiers trained from scratch can also achieve high performance for recognizing PPE items.

Additionally, the CNN classifiers with different depths have shown similar testing performances due to the relatively low resolution of cropped image patches (usually 32 × 32 or 64 × 64 in pixels) in this study. Meanwhile, the part attention regions retain the informative regions while eliminating distracting backgrounds, which also makes it easier for classifiers to extract critical features of PPE instances.

Fig. 16 shows typical mislabeled examples in the testing subset. In general, two reasons can explain these errors: (1) the low resolution of the input image patches and (2) visual confusion caused by similar objects. The CNN classifiers have difficulty in extracting sufficient features from low-resolution inputs. The alternative is to enhance image resolution with super-resolution (SR) techniques [53] before feeding them into the image classifiers. Similar objects can also lead to false-positive errors. For example, ordinary hats can be misleading objects for hardhats for their similarity in shapes. Introducing more negative examples during the training process could mitigate these false-positive errors.

5.4. Overall performance evaluation and comparative studies

In previous sections, the authors reported the performance of the three components that support the framework individually. This section provides the overall performance of the proposed method. An efficient PPE detection algorithm provides accurate classification results and localizes their classes with high IoU. *TP* is the number of correct PPE use detections with an IoU $\geq$ 0.5. *FP* is the number of PPE use detections with an IoU $<$ 0.5, while FN is the number of the PPE use instances that are not detected.

The test results in Table 7 show that the proposed framework yields a

Fig. 15. Examples of localization errors for part attention regions.

Table 5
Summary of the hardhat recognition results with different classifiers.

Task	Input size	Algorithms	Precision	Recall	F1-score	Speed (fps)
Hardhat recognition	32×32	VGG-16	97.52	**98.62**	98.07	104.9
		VGG-19	97.76	98.49	**98.12**	96.2
		MobileNet	94.92	96.11	95.51	49.3
		ResNet-18	97.35	96.61	96.98	59.3
		ResNet-34	97.59	96.49	97.04	37.3
		ResNet-101	**97.98**	97.62	97.80	15.8
		ResNet-152	97.82	95.86	96.83	10.5
		Proposed method	97.13	97.74	97.43	**179.3**

Note: precision, recall, and F1-score values are in percentage. The bold value indicates the highest performance.

Table 6
Summary of the vest recognition results with different CNN classifiers.

Task	Input size	Algorithm	Precision	Recall	F1-score	Speed (fps)
Vest recognition	64×64	VGG-16	**96.23**	97.53	**96.88**	99.9
		VGG-19	93.38	**98.20**	95.73	92.8
		MobileNet	94.35	86.29	90.14	48.3
		ResNet-18	94.95	93.03	93.98	55.8
		ResNet-34	89.44	89.44	89.44	37.5
		ResNet-101	92.34	89.44	90.87	15.5
		ResNet-152	93.94	90.56	92.21	10.4
		Proposed method	96.12	94.61	95.36	**156.6**

Note: precision, recall, and F1-score values are in percentage. The bold value indicates the highest performance.

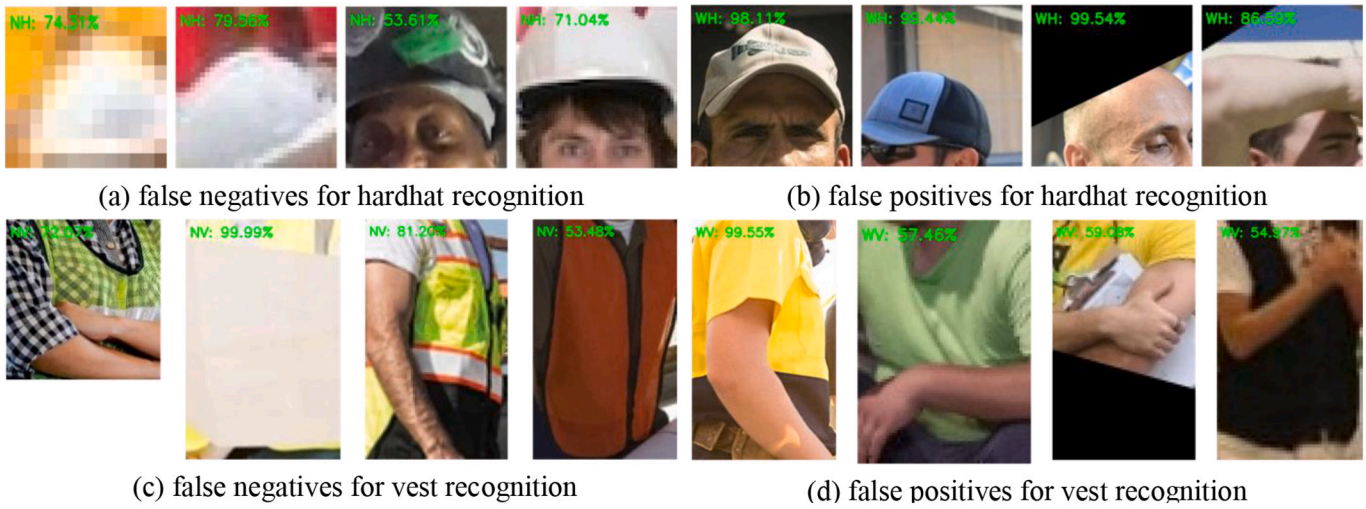


Fig. 16. Examples of mislabeled image patches.

Table 7
Detection results on the CPPE dataset.

Algorithm	Input size	Backbone	Hardhat detection			Vest detection		
			Precision	Recall	F1-score	Precision	Recall	F1-score
YOLO-v3 [54]	416×416	Darknet-53	94.70	90.58	92.59	92.10	67.63	77.99
SSD300 [19]	300×300	VGG-16	95.68	52.18	67.53	92.01	84.82	88.27
Faster R-CNN [20]	300×500	ResNet-50	79.01	89.99	84.14	90.76	92.97	91.85
Proposed method	32×32 / 64×64	–	**97.01**	**96.89**	**96.95**	**95.68**	**93.56**	**94.61**

Note: precision, recall, and F1-score values are in percentage. The bold value indicates the highest performance.

97.01% precision and a 96.89% recall for hardhat detection, and a 95.68% precision and 93.56% recall for vest detection. The performance of vest detection is slightly lower than hardhat recognition. Several factors may contribute to this issue. First, the upper body regions are more likely to be occluded by other workers, materials, or equipment in the workplace. Body occlusions can lead to additional distracting information (e.g., the partial bodies of other workers or site objects), even when the corresponding part attention regions are correctly localized and cropped. Second, particular orientations of workers (e.g., side-view) with fewer visual features of upper-body regions increase the task difficulty of vest recognition. Finally, distinguishing vests from ordinary clothing is challenging due to their visual similarity in textures, colors, and shapes. Fig. 17 shows the qualitative results of the proposed method. Different PPE classes are labeled in different colors to achieve

Fig. 17. Qualitative results of the proposed method. Different object categories are labeled in different colors. Red color – non-hardhat use; Green color – hardhat use; Yellow color – non-vest use; Blue color – vest use. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

better visualization.

The authors also compared the proposed method with current state-of-the-art methods [17,19,54] on the developed CPPE dataset. This study annotated all workers and PPE instances (e.g., hardhats and vests) in the CPPE dataset using LabelImg [55] as the Pascal VOC format. Ten percent of the images randomly selected from the training subset serve as validation subsets. To achieve better model performance, the authors first pre-trained all three models on the MS COCO dataset [51] and then fine-tuned the models on the CPPE dataset with a batch size of 8 using the Adam optimizer and an initial learning rate of 1e-4 for 100 epochs. The learning rate dropped by half every ten epochs. Table 7 reports the detection results of different methods on the CPPE testing subset.

As listed in Table 7, the proposed method provides higher precision and recall performance than existing methods. Specifically, the adopted detecting strategy surpasses the state-of-the-art methods by 1.33% precision (SSSD300) and 6.31% recall (YOLO-v3) in hardhat detection, as well as 3.58% precision (YOLO-v3) and 0.59% recall (Faster R-CNN) in vest detection. The Faster R-CNN model offers a relatively balanced performance for hardhat detection and vest detection. The SSD model commonly fails to detect hardhats, while the YOLO-v3 model yields relatively low recall performance on vest detection.

Fig. 18 shows a qualitative comparison of these approaches. The proposed method brings detection improvements by only focusing on local areas expecting PPE instances. Meanwhile, the developed strategy can directly determine non-PPE-use cases without any computational process of relationship verification of the cases involved. The skeleton-

Fig. 18. Qualitative comparisons with state-of-the-art methods. Examples annotated with white ovals correspond to false detections.

based representations of workers also help isolate each worker from a crowded workspace where severe occlusions exist between workers, compared to the worker detection in the form of bounding boxes. Finally, the pose-guided framework of this study supports the advantageous extensibility of detecting multi-class PPE items. The available keypoints can provide spatial anchors for localizing anticipated body part regions depending on PPE classes. When introducing a new PPE class, the classifiers trained to verify existing PPE types do not require any further re-training of the entire model.

5.5. Performance evaluation under challenging site scenarios

In this section, the authors qualitatively evaluated the performance of the proposed method under challenging site conditions, including tiny targets, extreme occlusions, non-regular illumination, low light, and blur, with the model parameters trained on the developed CPPE dataset. This study integrated the Pictor-v3 dataset [3] captured from construction sites as the additional testing dataset for these challenging scenarios. Since the Pictor-v3 dataset contains few safety vest-use examples, this study evaluated the model performance on worker and hardhat detection.

Table 8 shows the detection results of the YOLO-v3 model and the proposed method on the Pictor-v3 dataset. The results showed that the proposed method provides higher detection performance of workers and hardhats than the YOLO-v3 model. The challenging conditions mainly result in missing detections rather than false alarms. One of the most significant factors that could explain recall performance drop is that the Pictor-v3 dataset contained many tiny workers. As the tiny objects contain only a few pixels, the computer vision algorithms struggle to identify small-scale objects from long-range views [56]. Future research will examine optimal placements of cameras at construction sites to fully cover workplaces and improve image resolutions. In terms of enhancing algorithms for small object detection, recent model architectures such as Feature Pyramid Network (FPN) [57] that utilize multiscale features have shown promising performance for small object detection. The authors will explore the effectiveness of these two strategies for improving the performance of the developed approach in small object detection in future research.

Fig. 19 shows a qualitative comparison of these approaches. Since the object-centric models verify non-hardhat-use workers by checking whether a hardhat is present in or around a worker's detection region, the YOLO-v3 based detection scheme could generate false alarms while processing images having severe occlusions for human heads. However, the proposed method can significantly reduce the number of false alarms when body parts are invisible in the images. The qualitative results also demonstrated that other challenging conditions, such as low light, irregular illumination, or blur, did not significantly affect the performance of the proposed method in detecting workers and hardhats.

6. Limitations and discussions

The proposed method has shown several limitations in testing results for future improvements. First, the authors trained the pose estimation model on the publicly available MS COCO keypoint dataset [51]. The MS COCO keypoint dataset mainly collects images from daily life scenarios while containing few image samples from construction sites. Although the pre-trained model on the MS COCO dataset has achieved high worker detection performance in most cases, further efforts to establish a domain-specific dataset will enhance the model adaptation to pose estimation for construction workers. Second, the cropped body part attention regions are typically in low resolution (usually 32×32 or 64×64 in pixels). Such low-resolution examples are challenging for the classifiers to extract explicit features. The authors will integrate super-resolution (SR) techniques to improve the resolution of image patches before feeding them into the CNN classifiers. Third, to speed up the inference process, the authors used a lightweight MobileNet [44] with fewer parameters as the backbone of the pose estimation model and adopted a shallow CNN classifier rather than deep networks for PPE recognition. Simplifying the pose estimation model based on PPE types could further reduce computational resource requirements. For example, if the task were to identify the non-hardhat use workers, only the body joints in the head attention regions would need to be detected in images. Likewise, if the goal were to localize both non-hardhat use and non-vest use workers, lower body joints like ankles and knees are not necessarily required for this detection purpose. Furthermore, although the proposed method is extensible for verifying multi-class PPE compliance, the authors only focus on demonstrating the effectiveness of the proposed framework by simultaneously testing for safety violations of two types of PPE – hardhat and vest. The authors plan to extend the CPPE dataset to detect more types of PPE components, such as safety-toe footwear, gloves, or goggles.

Table 8
Detection results on the Pictor-v3 dataset.

Method	Worker detection			Hardhat detection		
	Precision (%)	Recall (%)	F1-score	Precision (%)	Recall (%)	F1-score
YOLO-v3 [54]	97.61	77.94	86.76	93.48	66.85	77.95
Proposed method	98.16	78.80	87.42	96.43	67.21	79.21

Note: precision, recall, and F1-score values are in percentage. The authors report the detection results by eliminating over-similar samples from the dataset to avoid evaluation bias.

7. Conclusions

Automatic monitoring for PPE use is crucial for ensuring safety controls and preventive measures at construction sites. This paper proposes a pose-guided anchoring framework to address the challenges of multi-class PPE detection in workspaces. The pose estimator first detects and represents individual workers in the form of full-body skeletons in crowded workplaces. The spatial anchors of worker poses can guide the algorithm's attention to specific body part attention regions that are anticipating PPE instances. The part attention localization module then integrated body knowledge-based rules to localize local image patches considering workers' orientations and object scales. This new strategy demonstrates its effectiveness in reducing search spaces while improving object recognition performance for handling multi-type PPE. Finally, this research trained two CNN-based classifiers to determine whether the identified part attention regions have hardhats or vests. Instead of verifying non-PPE use cases by checking spatial relationships of the involved workers and PPE instances, the new method directly inferred non-PPE cases from those regions where the expected PPE is missing.

To assess the performance of the proposed method, the authors established a new CPPE dataset of 932 images amounting to 2747 instances of hardhats, 1339 instances of safety vests, and 3428 workers. The experimental results on the developed CPPE dataset show that this new approach has achieved high precision and recall in individual tasks, i.e., worker detection, hardhat detection, and vest detection. Compared with the existing methods, the proposed method shows higher precision and recall performance in worker detection and PPE recognition. The pose-guided strategy also supports the advantageous extensibility of detecting multi-class PPE items. To encourage future research in the area, the authors have publicly released all trained models and the CPPE dataset in this paper on the GitHub page https://github.com/ruoxinx/PPE-Detection-Pose.

Nevertheless, the proposed framework has shown several limitations, and the authors also suggest possible directions for further improvements. First, establishing a construction domain-specific dataset

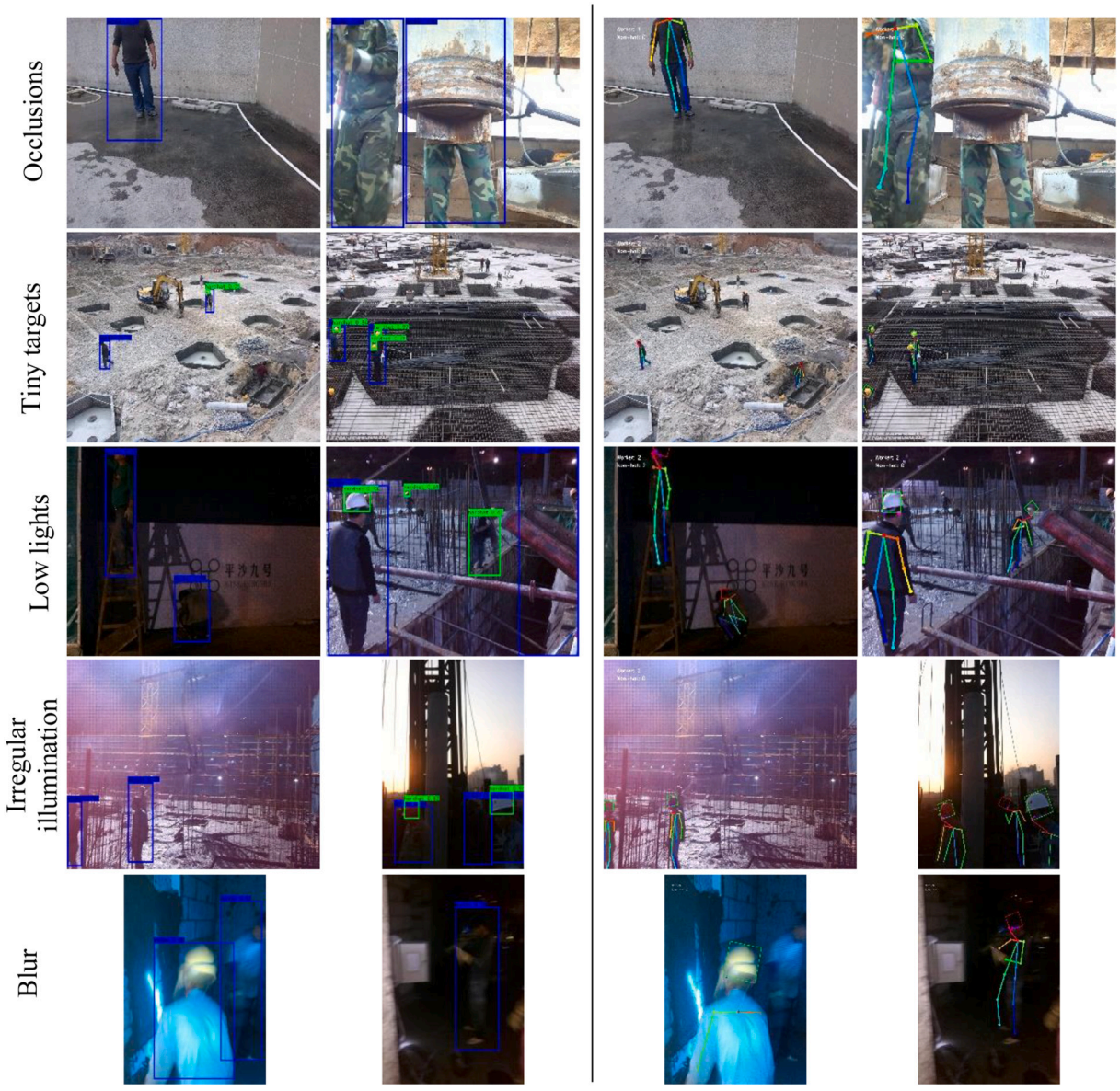


Fig. 19. Qualitative comparisons with state-of-the-art methods in challenging scenarios. **Left:** YOLO-v3 model; **Right:** The proposed method.

for worker pose estimation could enhance the model adaptation and reduce incorrect and missing detections. Second, the authors plan to simplify the pose prediction network based on the nature of the specific PPE detection task, which should further reduce the computation resource needs of the algorithm. Finally, the authors plan to extend the CPPE dataset to include more diverse PPE components.

Declaration of Competing Interest

None.

Acknowledgments

This material is based on work supported by the U.S. National Science Foundation (NSF) under Grant No. 1454654 and 1937115, the Nuclear Energy University Program (NEUP) of the U.S. Department of Energy (DOE) under Award No. DE-NE0008864, and College of Engineering Dean's Fellowship of the Carnegie Mellon University (CMU). The supports are gratefully acknowledged. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF, DOE, and CMU.

References

[1] Bureau of Labor Statistics (BLS), Number and rate of fatal work injuries (by industry). https://www.bls.gov/charts/census-of-fatal-occupational-injuries/number-andrate-of-fatal-work-injuries-by-industry.htm. Accessed date: 3 April, 2020.
[2] F. Akbar-Khanzadeh, Factors contributing to discomfort or dissatisfaction as a result of wearing personal protective equipment, Journal of Human Ergology 27 (1998) 70–75, https://doi.org/10.11183/jhe1972.27.70.
[3] N.D. Nath, A.H. Behzadan, S.G. Paal, Deep learning for site safety: real-time detection of personal protective equipment, Autom. Constr. 112 (2020) 103085, https://doi.org/10.1016/j.autcon.2020.103085.
[4] J. Shen, X. Xiong, Y. Li, W. He, P. Li, X. Zheng, Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning, Computer-Aided Civil and Infrastructure Engineering (2020) 1–17, https://doi.org/10.1111/mice.12579.
[5] M.W. Park, N. Elsafty, Z. Zhu, Hardhat-wearing detection for enhancing on-site safety of construction workers, J. Constr. Eng. Manag. 141 (9) (2015), 04015024, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000974.
[6] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, C. Li, Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment, Autom. Constr. 93 (2018) 148–164, https://doi.org/10.1016/j.autcon.2018.05.022.

[7] The Occupational Safety and Health Administration (OSHA), Personal protective equipment. https://www.osha.gov/Publications/osha3151.pdf. Accessed date: 3 April, 2020.
[8] Safety and Health Magazine, OSHA's top 10 most cited violations for 2019. https://www.safetyandhealthmagazine.com/articles/19087-oshas-top-10-most-cited-violations?utm-source=from-news-brief&utm-campaign=top10.
[9] Oregon Occupational Safety and Health Administration, Personal protective equipment–Oregon OSHA online course 1241. https://ehs.oregonstate.edu/sites/ehs.oregonstate.edu/files/pdf/occsafety/or-osha_ppe_training.pdf. Accessed date: 3 April, 2020.
[10] OSHA, OSHA penalties. https://www.osha.gov/laws-regs/oshact/section_17. Accessed date: 3 April, 2020.
[11] A. Kelm, L. Laußat, A. Meins-Becker, D. Platz, M.J. Khazaee, A.M. Costin, et al., Mobile passive radio frequency identification (RFID) portal for automated and rapid control of personal protective equipment (PPE) on construction sites, Autom. Constr. 36 (2013) 38–52, https://doi.org/10.1016/j.autcon.2013.08.009.
[12] H. Li, X. Li, X. Luo, J. Siebert, Investigation of the causality patterns of non-helmet use behavior of construction workers, Autom. Constr. 80 (2017) 95–103, https://doi.org/10.1016/j.autcon.2017.02.006.
[13] S.H. Kim, C. Wang, S.D. Min, S.H. Lee, Safety helmet wearing management system for construction workers using three-axis accelerometer sensor, Appl. Sci. 8 (2018) 2400, https://doi.org/10.3390/app8122400.
[14] H. Wu, J. Zhao, An intelligent vision-based approach for helmet identification for work safety, Comput. Ind. 100 (2018) 267–277, https://doi.org/10.1016/j.compind.2018.03.037.
[15] B.E. Mneymneh, M. Abbas, H. Khoury, Vision-based framework for intelligent monitoring of hardhat wearing on construction sites, J. Comput. Civ. Eng. 33 (2) (2019), 04018066, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000813.
[16] R. Girshick, Fast R-CNN, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448, https://doi.org/10.1109/ICCV.2015.169.
[17] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.
[18] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.
[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, et al., SSD: Single shot multibox detector, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 21–37, https://doi.org/10.1007/978-3-319-46448-0_2.
[20] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T.M. Rose, et al., Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Autom. Constr. 85 (2018) 1–9, https://doi.org/10.1016/j.autcon.2017.09.018.
[21] W. Fang, L. Ding, H. Luo, P. Love, Falls from heights: a computer vision-based approach for safety harness detection, Autom. Constr. 91 (2018) 53–61, https://doi.org/10.1016/j.autcon.2018.02.018.
[22] J. Wu, N. Cai, W. Chen, H. Wang, G. Wang, Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset, Autom. Constr. 106 (2019) 102894, https://doi.org/10.1016/j.autcon.2019.102894.
[23] S. Tang, D. Roberts, M. Golparvar-Fard, Human-object interaction recognition for automatic construction site safety inspection, Autom. Constr. 120 (2020) 103356, https://doi.org/10.1016/j.autcon.2020.103356.
[24] S. Chen, K. Demachi, A vision-based approach for ensuring proper use of personal protective equipment (PPE) in decommissioning of Fukushima Daiichi nuclear power station, Appl. Sci. 10 (2020) 5129, https://doi.org/10.3390/app10155129.
[25] S. Du, M. Shehata, W. Badawy, Hard hat detection in video sequences based on face features, motion and color information, in: Proceedings of the International Conference on Computer Research and Development, 2011, pp. 25–29, https://doi.org/10.1109/ICCRD.2011.5763846.
[26] K. Shrestha, P.P. Shrestha, D. Bajracharya, E.A. Yfantis, Hard-hat detection for construction safety visualization, Journal of Construction Engineering (2015) 721380, https://doi.org/10.1155/2015/721380.
[27] A. Yao, J. Gall, G. Fanelli, L. Van Gool, Does human action recognition benefit from pose estimation? Proceedings of the British Machine Vision Conference, 2011, https://doi.org/10.5244/C.25.67, 67.1–67.11.
[28] K. Yuen, M.M. Trivedi, Looking at hands in autonomous vehicles: a convnet approach using part affinity fields, IEEE Transactions on Intelligent Vehicles 5 (3) (2019) 361–371, https://doi.org/10.1109/TIV.2019.2955369.
[29] J. Charles, T. Pfister, M. Everingham, A. Zisserman, Automatic and efficient human pose estimation for sign language videos, Int. J. Comput. Vis. 110 (2014) 70–90, https://doi.org/10.1007/s11263-013-0672-6.
[30] S. Yeung, F. Rinaldo, J. Jopling, B. Liu, R. Mehra, N.L. Downing, , et al.A. Milstein, A computer vision system for deep learning-based detection of patient mobilization activities in the ICU, NPJ Digital Medicine 2 (2019) 11, https://doi.org/10.1038/s41746-019-0087-z.
[31] M.A. Fischler, R.A. Elschlager, The representation and matching of pictorial structures, in: IEEE Transactions on Computers, C–22 1, 1973, pp. 67–92, https://doi.org/10.1109/T-C.1973.223602.
[32] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1385–1392, https://doi.org/10.1109/CVPR.2011.5995741.
[33] A. Toshev, C. Szegedy, DeepPose: Human Pose Estimation Via Deep Neural Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660, https://doi.org/10.1109/CVPR.2014.214.
[34] S.E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional Pose Machines, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732, https://doi.org/10.1109/CVPR.2016.511.
[35] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 483–499, https://doi.org/10.1007/978-3-319-46484-8_29.
[36] B. Sapp, B. Taskar, MODEC: Multimodal decomposable models for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3674–3681, https://doi.org/10.1109/CVPR.2013.471.
[37] Z. Cao, T. Simon, S.E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1302–1310, https://doi.org/10.1109/CVPR.2017.143.
[38] Z. Sun, C. Zhang, J. Chen, P. Tang, A. Yilmaz, Predictive nuclear power plant outage control through computer vision and data-driven simulation, Prog. Nucl. Energy 127 (2020) 103448, https://doi.org/10.1016/j.pnucene.2020.103448.
[39] D. Roberts, W. Torres Calderon, S. Tang, M. Golparvar-Fard, Vision-based construction worker activity analysis informed by body posture, J. Comput. Civ. Eng. 34 (4) (2020), 04020017, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000898.
[40] S.J. Ray, J. Teizer, Real-time construction worker posture analysis for ergonomics training, Adv. Eng. Inform. 26 (2) (2012) 439–455, https://doi.org/10.1016/j.aei.2012.02.011.
[41] A. Peddi, L. Huan, Y. Bai, S. Kim, Development of Human Pose Analyzing Algorithms for the Determination of Construction Productivity in Real-Time, Construction Research Congress, Seattle, WA, 2009, pp. 11–20, https://doi.org/10.1061/41020(339)2.
[42] M. Liu, S. Han, S. Lee, Potential of convolutional neural network-based 2D human pose estimation for on-site activity analysis of construction workers, in: Proceedings of the Computing in Civil Engineering, 2017, pp. 141–149, https://doi.org/10.1061/9780784480847.018.
[43] X. Yan, H. Li, C. Wang, J. Seo, H. Zhang, H. Wang, Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion, Adv. Eng. Inform. 34 (2017) 152–163, https://doi.org/10.1016/j.aei.2017.11.001.
[44] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., MobileNets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint (2017). https://arxiv.org/abs/1704.04861.
[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint (2014). https://arxiv.org/abs/1409.1556.
[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9, https://doi.org/10.1109/CVPR.2015.7298594.
[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y.
[49] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (2015) 2278–2324, https://doi.org/10.1109/5.726791.
[50] GitHub, Safety helmet (hardhat) wearing detect dataset. https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset, 2019. Accessed date: 3 April, 2020.
[51] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: Common objects in context, Proceedings of the European Conference on Computer Vision, 2014, pp. 740–755, https://doi.org/10.1007/978-3-319-10602-1_48.
[52] GitHub, roLabelImg. https://github.com/cgvict/roLabelImg. Accessed date: 3 April, 2020.
[53] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 38 (2) (2016) 295–307, https://doi.org/10.1109/TPAMI.2015.2439281.
[54] J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, arXiv preprint, arXiv:1804.02767, arXiv preprint, 2018, https://arxiv.org/abs/1804.02767.
[55] GitHub, LabelImg: a Graphical Image Annotation Tool. https://github.com/tzutalin/labelImg. Accessed date: 3 April, 2020.
[56] N.D. Nath, A.H. Behzadan, Deep convolutional networks for construction object detection under different visual conditions, Front. Built Environ. 6 (2020) 97, https://doi.org/10.3389/fbuil.2020.00097.
[57] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944, https://doi.org/10.1109/CVPR.2017.106.