Automation in Construction 176 (2025) 106231

Contents lists available at ScienceDirect

Automation in Construction

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

Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation

Leopoldo López [a], Jonay Suárez-Ramírez [a], Miguel Alemán-Flores [b], Nelson Monzón [b,*]

[a] Qualitas Artificial Intelligence and Science, Parque Científico Tecnológico, 35017, Las Palmas de Gran Canaria, Spain
[b] Instituto Universitario de Cibernética, Empresa y Sociedad (IUCES), University of Las Palmas de Gran Canaria (ULPGC), Campus Universitario de Tafira, 35017, Las Palmas de Gran Canaria, Spain

ARTICLE INFO

Keywords:
PPE compliance
Deep learning
Worker detection
Pose estimation
Health
Safety and environment (HSE)

ABSTRACT

This paper presents an AI framework for automated detection of personal protective equipment (PPE) compliance in complex construction and industrial environments. Ensuring health and safety standards is essential for protecting workers engaged in construction, repair, or inspection activities. The framework leverages deep learning techniques for worker detection and pose estimation to enable accurate PPE identification under challenging conditions. The framework components are replaceable, and employ the InternImage-L detector for worker detection, ViTPose for pose estimation, and YOLOv7 for PPE recognition. A duplicate removal stage, combined with pose information, ensures PPE items are accurately assigned to individual workers. The approach addresses challenges like shadows, partial occlusions, or densely grouped workers. Evaluated on diverse datasets from real-world industrial settings, the framework achieves competitive precision and recall, particularly for critical PPE like helmets and vests, demonstrating robustness for safety monitoring and proactive risk management.

1. Introduction

Construction sites are high-risk environments where workers are exposed to substantial occupational hazards [1]. Adherence to health, safety, and environmental (HSE) regulations is essential to mitigate these risks and prevent accidents. Personal protective equipment (PPE) is among the most effective measures to reduce injuries, since many accidents, such as falls, electrocution, and being struck by objects, can be prevented through proper compliance with protective equipment [2]. Continuous monitoring of worker behavior and work conditions is critical to identify hazards and enforce compliance with safety protocols. Traditional human supervision is often insufficient in construction sites due to their highly dynamic environments, characterized by frequent interactions between machinery and workers.

Ensuring worker safety requires not only the use of PPE, but also the implementation of appropriate tools and systems to monitor and enforce safety protocols [3]. Construction sites, especially outdoor scenarios, present significant challenges, such as occlusions, overlapped workers, shadowed areas or variable luminance. Current PPE detection methods, particularly those based on neural networks, have demonstrated promising results, but limitations still remain under these challenging conditions [4,5], specifically when detecting small PPE items or distant workers, as well as in cluttered environments with frequent worker and equipment overlaps. Additionally, variable lighting conditions and occlusions introduce further difficulties, in particular when using single-camera systems.

Consequently, the ongoing need for robust systems capable of reliably identifying workers and accurately assigning PPE to the correct individuals continues to attract researchers to contribute to the literature. In this sense, our paper aims to develop a robust system for real-time monitoring of PPE compliance using advanced computer vision and deep learning techniques.

Our research originates from the SIVIS initiative,[1] which aims to enhance safety and operational efficiency in shipyards through the application of artificial intelligence (AI) techniques. Although this project primarily focuses on workers in shipyards, the environment and activities share significant similarities with other construction sites, where the risks and PPE requirements are quite similar. Our current work extends to those additional scenarios, addressing safety concerns in various high-risk environments.

We present a framework designed to collectively address the limitations mentioned above. Unlike some previous approaches, where the pose estimation is used as a tool for restricting PPE search areas, our strategy detects PPE independently and then leverages pose

* Corresponding author.
E-mail address: nelson.monzon@ulpgc.es (N. Monzón).
[1] http://sivisrepnaval.com/.

https://doi.org/10.1016/j.autcon.2025.106231
Received 12 November 2024; Received in revised form 14 April 2025; Accepted 22 April 2025
Available online 8 May 2025
0926-5805/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

information to refine worker-PPE matching. Our approach avoids the dependency on perfect pose estimation, reducing errors when body keypoints are inaccurately detected. Additionally, we introduce a collision detection and duplicate removal module, which mitigates errors caused by overlapping workers—an issue that remains a challenge for many state-of-the-art methods. Supported by these components, our framework provides robust results for multiple challenges, particularly in complex scenes with high worker density.

Our proposal leverages advanced computer vision and deep learning techniques to monitor PPE compliance in real-time. Workers are detected using a state-of-the-art model and cropped images of individuals are processed to identify PPE items and generate skeletal poses for precise PPE-to-body-part alignment. To enhance the visibility of PPE, linear interpolation is applied to resized crops, facilitating a better feature extraction for subsequent detections. We introduce mechanisms for collision detection and duplicate removal to resolve ambiguities caused by overlapping workers and improve the reliability of PPE assignment. The modular design ensures flexibility, allowing each component to be replaced with alternatives without compromising system performance.

Pose estimation has previously been used, but mainly to search for PPE items around the corresponding body part, instead of detecting them independently and using pose information and distances for the matching process (e.g., Vukicevic et al. [6] or Xiong et al. [7]). Some other works have improved PPE-worker association, but struggle with overlapping workers which make the matching worker-PPE particularly difficult (e.g., Nath et al. [8]). In this regard, we have introduced a collision detection and duplicate removal stage to ease this association. Finally, some previous works rely on multi-camera re-identification, but do not address single-camera scenarios, where occlusions are difficult to tackle (e.g., Cheng et al. [9]). Our proposal is built on similar concepts, while introducing the combination of pose estimation, collision detection, and duplicate removal pursuing a reliable PPE assignment even in complex scenarios.

In our experiments, we validate the framework by performing an extensive evaluation comparing it with an improved version of [8], and our own implementations of [7]. These approaches have been chosen due to their relevance in PPE compliance monitoring. Additionally, we assess the modularity of our framework by testing different configurations of its components, quantitatively evaluating the impact of alternative worker detectors and pose estimation.

Experiments on datasets from real construction sites and the Repnaval shipyard demonstrate high precision and recall, particularly for critical PPE items such as helmets and vests. By addressing key limitations and introducing mechanisms to enhance robustness and flexibility, this work demonstrates significant improvements in monitoring PPE compliance in real-world construction and shipyard environments. The performance of our approach is also analyzed under various environmental conditions, worker densities, and PPE compliance scenarios, to validate its applicability in diverse real-world settings and to further assess its generalization capabilities.

The remainder of this paper is structured as follows. Section 2 reviews relevant studies on PPE detection, worker identification, and pose estimation, highlighting the limitations of existing approaches. Section 3 describes the proposed framework, detailing its modular design. Section 4 presents the experimental setup, including dataset details, preprocessing steps, and evaluation metrics. Section 5 analyzes the quantitative and qualitative results, comparing our approach with state-of-the-art methods and assessing its performance under different challenging conditions. Additionally, we evaluate the modularity of our framework by testing different configurations of its components and analyzing their impact on performance. An ablation study is also presented. Finally, Section 6 summarizes the main conclusions of our work.

2. Related works

Several works have addressed the automated detection of personal protective equipment (PPE) and worker pose, proposing different perspectives. Detecting them is especially challenging in open spaces, where factors such as changing illumination, large distances from the camera, shadows and occlusions come into play.

The detection of small objects on construction sites for safety reasons has been addressed in works such as [10], where a small object detection (SOD) system is presented for comprehensive site monitoring, based on the YOLOv5 algorithm. This work applies a multiscale approach to deal with objects of different sizes. Small objects here refer to those viewed from medium to large distances, appearing small in the image. This is often the case in the scenes analyzed in this work.

Regarding the identification of PPE, most works focus on specific elements of the attire to identify them based on their features. Thus, a large number of vision-based approaches for monitoring PPE compliance focus on identifying hard hats. For instance, the work by Mneymneh et al. [11] evaluates the performance of feature detection, extraction and matching; template matching; and cascade classifiers for the detection of hard hats in construction environments. Fang et al. [12] use region-based convolutional neural networks (R-CNNs) to detect if a worker is wearing a hard hat, while Xie et al. [13] use fully convolution-based algorithms to detect hard hats. In [14], the authors also focus on hard hats and identify whether workers on construction sites wear them and what color they are. They propose a one-stage CNN-based method which aggregates multiscale features to deal with small-scale hard hats. Wu et al. [15] propose a semantic attribute recognition based on the transformer architecture that identifies whether workers are wearing a helmet and a harness. Yang et al. [16] also focus on helmets and have created a dataset with images of people working in construction sites.

Although hats are the most commonly detected items, several works analyze other components of the attire. For instance, Wang et al. [17] deal with the detection of helmets in four colors, but also include safety vests and the detection of people themselves. A graph neural network is used by Zhao et al. [18] to detect four types of hats, vests, glasses, as well as people and their heads. This work performs an arc-flash analysis to assess the electrical hazard using a data augmentation few-shot approach. In certain cases, additional information is utilized to facilitate the identification of the attire. For example, Seong et al. [19] consider the particular color of protection vests to identify them and then extract the workers in the scene. Nevertheless, these restricted approaches limit the extension, scalability and broader applicability of the method.

Some works address challenging conditions, such as poor illumination and small object sizes. For instance, Wang et al. [20] include a large-size input layer for multi-scale prediction and adjust the size of anchor boxes to cope with small helmets and protective clothes. In [21], the authors apply a neural net scheme under poor illumination conditions. The work in [22] focuses on the detection of hard hats, in both images and video, using color as well as monochrome images.

Methods can also be categorized into those using single shot detectors (SSD), such as [14] for hard hat detection, and those processing a whole video sequence, such as [8]. The work by Cheng et al. [9] combines worker re-identification and PPE classification. A loss function is included to learn more discriminative human features and improve the tracking of individual workers. Furthermore, a weighted-class strategy is used to reduce the impact of imbalance among classes.

In [3], the authors isolate moving workers and any hard hat around the top area of a worker's box is located. The work in [15] extracts attention regions for the identification of helmets and harnesses and presents images with an increased complexity,[2] because of occlusions, overlapping and illumination problems.

[2] https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset.

Table 1
Summary of the most relevant works in PPE compliance monitoring and key aspects of each proposal.

Reference	Description
Xiong et al. [7]	Pose estimation to identify head and upper body. Binary classifier to determine if a worker is wearing a hat/vest.
Nath et al. [8]	Three approaches using YOLO-v3: (a) Independent worker and PPE detection, then linked. (b) Single object detection for workers and PPE. (c) Step 1: Worker detection, Step 2: PPE detection around them.
Cheng et al. [9]	Multiple cameras, moving workers. Worker re-identification. Weighted class strategy for PPE classification.
Vukicevic et al. [6]	18 different PPE classes and 5 body parts. Pose estimator used to define regions of interest. MobileNetV2 recommended.

The information extracted from the scenes is crucial to analyze the risks and identify dangerous situations. Thus, workers can be detected in images, tracked in video sequences, and analyzed for their safety state based on extracted features. In [23], a standard version of the YOLOv3 is applied to classify the scene into different categories according to the presence of safety attire. In [24], the authors present different object recognition models to identify workers, their risks and the use of appropriate safety elements.

The combination of the different PPE elements and the worker's pose is also important for certain analyses. In this way, several works combine the detection of PPE with identifying workers, their heads, or their poses. Vukicevic et al. [6] consider 18 PPE types and 5 body regions. The regions of interest are extracted using the HigherHRNet pose estimator and classified by means of different image classification architectures. Pose estimation is used to define the regions of interest prior to PPE detection. This strategy differs from ours, which uses pose estimation to assign the PPE items to the right worker using distances from PPE to related body parts. A similar approach is proposed in Xiong et al. [7], which also uses the location of the estimated body parts to search for the corresponding PPE items using specific classifiers for each body part. Although a prior location of body parts can restrict the area to look for the PPE, the dependency on a right pose estimation can lead to missing PPE items. Kim et al. [25] estimate the 3D pose from 2D images by using a single camera and a virtual model for the 2D–3D annotation. The work in [26] proposes a deep learning framework to extract the worker's keypoints, track multiple workers, and analyze their activity. In [27], the authors make use of decision trees to detect different helmets and gears and classify poses into several classes. Several approaches to pose estimation use non-visual systems, which mainly rely on the workers wearing sensors and collecting the position together with certain parameters, such speed, direction, etc. (see, for instance, [28]). Alternatively, depth cameras provide some valuable information for pose estimation [29]. However, both options require specialized equipment that is not commonly available or practical for use on construction sites or in shipyards. In addition, the limited shooting distance of depth cameras and the sensitivity to certain conditions restrict their use in outdoor scenes. A comprehensive analysis of safety conditions could also include the estimation of the pose of the construction equipment and machines, such as in [30]. A vision-based collision warning system is presented in [31], which warns of potential risks using an automated 3D position estimation of each worker with monocular vision.

Different versions of the YOLO (You Only Look Once) model are used in a significant proportion of these works. Nath et al. [8] present a comparison of three different strategies to analyze image and video sequences in real time and determine whether workers are wearing a hard hat or a vest. All three approaches rely on YOLO-v3, but differ in the way workers, hats and vests are detected. With the first method, the detection of a worker and their PPE is handled as two distinct tasks, and after both are detected, the system tries to associate the PPE with the corresponding worker. The second approach uses a single object detector to identify workers and PPE, and classifies workers according to the PPE they are wearing. Instead of detecting workers and PPE separately, the model identifies and categorizes them together in one step, enabling faster and more integrated detection. The third one first detects workers and then focuses on the areas around the workers (a cropped region of the image) to detect the PPE they are wearing. After identifying a worker, the system zooms into that region to classify the PPE using a multiclass classifier, distinguishing between different types of safety gear.

Some works perform comparisons between different models or add certain modifications to improve their efficacy. For instance, Akinsemoyin et al. [32] compare faster R-CNN and YOLOv3 when applied to images acquired with unmanned aerial systems in order to detect hard hats in construction sites. Nain et al. [33] tested the accuracy of three deep learning algorithms (YOLO v4, v5, and YOLACT++) in the detection of hard hats. Zhang et al. [34] perform certain modifications in the YOLOv5s-Btri model to improve the detection of helmets and reduce the parameters.

Concerning the datasets available for testing the different approaches, [22,32] use a publicly available dataset[3] to determine whether the workers are wearing a hard hat. The work in [18] copes with a wide variety of elements,[4] and Wang et al. [17] have made their dataset available[5] to compare their results in the identification of different elements of the attire. However, most of those images are not as challenging as the scenarios we deal with because of the proximity of the workers to the camera. The scenario most similar to ours is used in [8], with a dataset which includes images of construction sites with workers at a relatively greater distance from the camera.[6]

Table 1 summarizes the main features of some of the most relevant works, which will be used for comparison in the following sections. As mentioned above, Nath et al. [8] include three alternatives. In the first one, workers and PPE items are detected independently and linked afterwards. The second option performs a common object detection for workers and PPE. Finally, in the third alternative, workers are detected first and PPE items are searched for around them. The latter two approaches share some aspects with our proposal, but they lack mechanisms to successfully decide when the worker-PPE matching is ambiguous. We have included a collision detection and a duplicate removal stage that benefits our results in those cases. Additionally, pose estimation is used to improve the accuracy of the assignment of equipment to workers. On the other hand, Xiong et al. [7] and Vukicevic et al. [6] rely on pose estimation to search for the equipment around the corresponding body parts. This limits the possibilities of successfully finding PPE items. Finally, Cheng et al. [9] deal with several cameras and track workers with a re-identification process, whereas we work

[3] https://public.roboflow.com/object-detection/hard-hat-workers.
[4] https://github.com/msdbarati/PPE-Detection.
[5] https://github.com/ZijianWang-ZW/PPE_detection.
[6] https://universe.roboflow.com/ppe-orxtt/ppe-u7jtr/dataset/9.

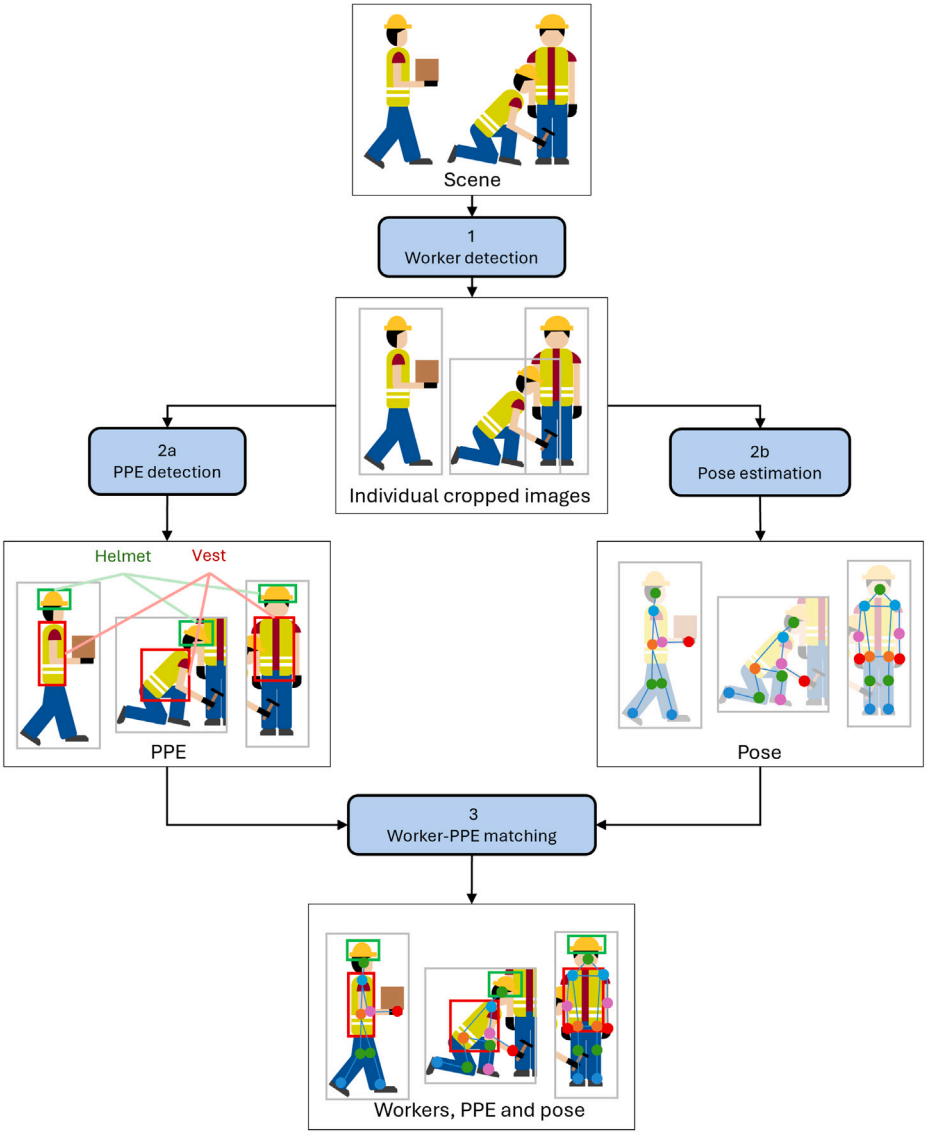

Fig. 1. General flow diagram for the detection and matching of workers and equipment.

with single-camera systems. With the additional stages that we have introduced and the new strategy in the search and assignment of PPE items, the results are more precise and trustworthy in those challenging situations where the detection and the correct matching are especially difficult.

3. Proposed framework for worker and equipment detection

In order to provide useful information for assessing the risks associated with a scene, it is essential not only to detect workers and different items of PPE, but also to match the detected equipment with the corresponding worker. This process involves addressing challenging scenarios which include very close workers, ambiguities in the assignment of a PPE item to a worker, and partial occlusions.

Fig. 1 presents a general flow diagram that illustrates the different stages, as well as their inputs and outputs. First, workers are detected in the original scene, and cropped images are extracted around each worker for further analysis (step 1 in the diagram). Then each cropped image is processed in two parallel pathways. On the one hand, the different elements of the PPE are identified within the cropped image (step 2a). This includes helmets, vests, gloves, safety boots, and protective suits. Within each image, the different elements of the attire are labeled according to their categories (for clarity, only helmets and vests are depicted in the diagram). Since workers may be in close proximity and the corresponding cropped images may overlap, a duplicate removal stage is necessary to ensure that the detected items are not considered in more than one cropped image.

On the other hand, the pose of the worker is analyzed by extracting a skeletal representation of the position of several body parts (step 2b). Note that steps 2a and 2b are independent and can be performed in parallel. Finally, a matching process is performed to associate the detected workers with their respective equipment (step 3). Although the risk associated with the pose itself is not considered, it facilitates a more accurate matching by aligning PPE elements with the corresponding body parts.

This flexible framework is designed in such a way that each block can be replaced with an alternative that performs the same task, using the same inputs and outputs, but employing a different technique. In the following subsections, each stage will be described in greater detail.

3.1. Worker detection and generation of cropped images

For people detection (step 1 in Fig. 1), any object detector trained to identify individuals can be utilized (e.g., YOLO or InternImage). In this

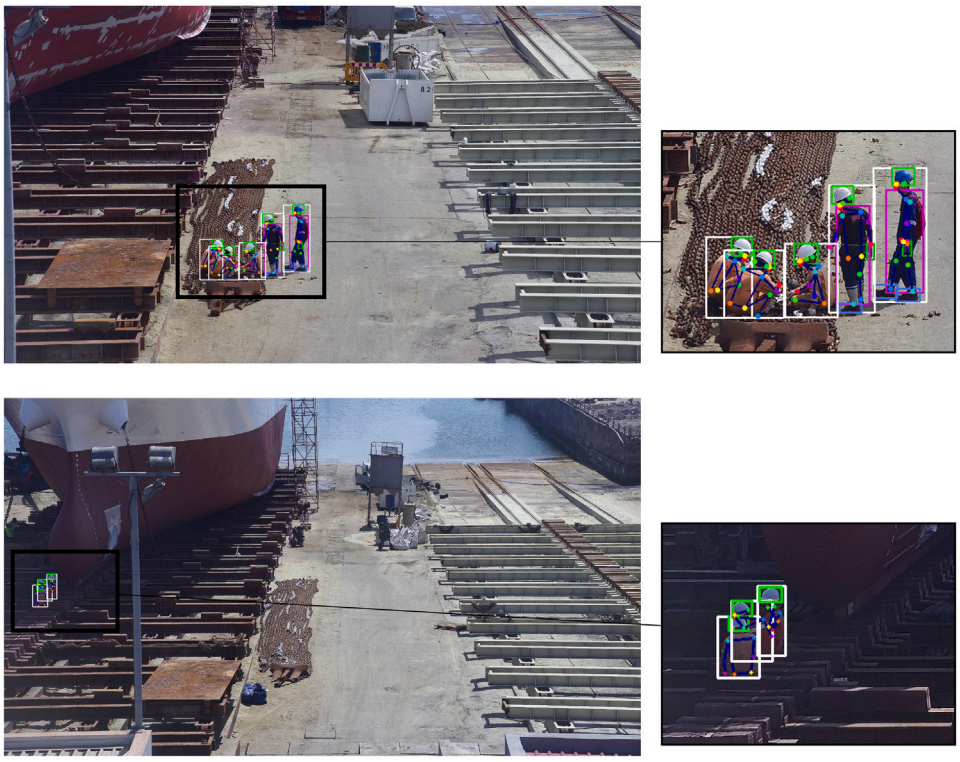

Fig. 2. Results for worker and equipment detection in large distance scenarios with shadows and occlusions. The following color code has been used to indicate the detection categories: worker helmet vest gloves safety boots protective suit.

study, we implemented the proposed framework using InternImage-L, trained on the COCO dataset. Our experiments indicate that this is the most suitable option due to its ability to capture long-range dependencies and adapt to complex spatial variations through its large-kernel convolution and deformable attention mechanisms.

As previously mentioned, for each detected worker, a section of the image is cropped around them, ensuring full inclusion. To optimize PPE detection, these crops are interpolated to a uniform size of 192×192 pixels. This process enhances the visibility of small or distant PPE items by increasing their resolution, making their features more distinguishable to the detection model. Additionally, in scenarios involving partial occlusions — whether between workers or with surrounding objects — interpolation helps to preserve visible details, improving the likelihood of detecting PPE components even when they are only partially visible. While interpolation does not introduce new information, it ensures that PPE elements remain within an optimal scale range for feature extraction, benefiting models that are sensitive to object size variations and occlusions.

When the aspect ratio does not match the target size, the remaining space is filled with gray pixels to maintain a consistent input format. Alternatively, zero-padding or edge replication can be used, depending on the model's preprocessing requirements.

The right identification of workers is crucial for the subsequent stages. A false negative in the detection of a worker results in a missing cropped image and overlooking the possible PPE items. On the other hand, a false positive (i.e., identifying a worker where no one is present) could lead to triggering a warning for not complying with the required regulations (a false alarm). Finally, overlapping and occlusions must be cautiously tackled to assign PPE items to the right workers. These aspects will be considered in the following sections.

3.2. Equipment identification with collision detection and duplicate removal

In the proposed framework, an object detector is used to extract PPE from each cropped image (step 2a in Fig. 1). Specifically, we employ YOLOv7 [35] and identify five categories: helmets, vests, gloves, safety boots, and protective suits.

This process may result in the detection of duplicates of some items, when they are visible in more than one subimage, thus requiring the removal of the duplicates. Fig. 2 illustrates these outcomes, where the cropped images generated for different workers overlap, and some workers are occluded by other workers, scaffolding, or other elements. The left side of the figure displays the original images alongside the inferred detection, while the right side provides a detailed view of regions where multiple workers and their equipment are detected. The top image depicts five workers (white rectangles) who are very close to each other. Notably, the two on the far left are particularly nearby, which may result in their helmets (green rectangles) being detected in more than one cropped image. A similar situation is observed with the workers in the bottom image, where shadows further complicate the identification process.

To tackle this issue, we apply non-maximum suppression (NMS) to the detected equipment across all cropped images, after transforming the detected elements back to their original dimensions and positions in the image. By utilizing the NMS technique, we filter the equipment detected in various subimages as follows: The coordinates of the instances detected in the different cropped images are recalculated with respect to the entire image. When two instances of the same type overlap beyond a certain threshold (i.e., their Intersection over Union, or IoU, is sufficiently high), only that with the highest confidence score (as provided by the detection model) is retained, while the others are discarded. In our experiments, this value is set to 0.65. This way, the attire contained in more than one cropped image (as in the lower row of Fig. 2) is not detected multiple times. As observed, the main issue regarding PPE identification is related to occlusions and shadows, which may result in missing PPE items or ambiguous ownership. However, the collision detection and duplicate removal stage helps dealing with overlapping cropped images and the pose estimation described below improves the efficacy of worker-PPE matching. Section 5 shows more experimental results when the method is applied to important challenges.

3.3. Pose extraction

Pose estimation is performed using ViTPose [36], specifically ViTPose-B trained on the COCO dataset (step 2b in Fig. 1). This method generates a skeletal representation of the workers consisting of edges that connect vertices, which correspond to eyes, ears, nose, shoulders, elbows, wrists, hips, knees, and ankles. It provides valuable information for accurately matching workers with their respective equipment, as described below. In our experiments, we quantitative confirm that ViTPose is the best option for our framework.

RRIt is important to note that pose is used to increase the accuracy of the matching between a worker and the PPE items they wear. However, in our proposal this matching is carried out even if pose is missing. In fact, in those cases where pose cannot be extracted, the PPE item is assigned to the worker whose cropped image contained the item. In case it overlaps with another cropped image, it is associated with the worker whose bounding box center is closer to the item. This approach differs from that in [6] or [7], which focus attention around pose keypoints and rely on pose estimation for a proper PPE identification.

3.4. Worker-equipment matching

Workers often collaborate on tasks in industrial scenarios, resulting in close proximity and frequent partial or total occlusions. In Section 3.2, we described the process of removing duplicate equipment detected across different cropped images. However, an additional process is required to accurately identify the worker who is wearing the PPE item. This challenge is addressed by integrating the PPE detection from each worker's subimage with their estimated pose (step 3 in Fig. 1).

Since each item of the PPE must be worn in a specific part of the body, pose information helps assign each item to the corresponding worker by identifying the body part where it should be located. By integrating this pose information with the detected PPE, we can effectively match each worker with their appropriate equipment and assess the compliance with HSE regulations in a more effective way. This step is illustrated in Fig. 3.

To assign ownership of each PPE item, our method calculates the distance from the center of the PPE item's bounding box to the keypoints of the corresponding body part, allowing for association with the closest relevant body part (e.g., a helmet with the head). This calculation is only performed for overlapping workers to optimize processing time. Subsequently, the PPE is assigned to the worker whose pose keypoints are nearest to the center of the item, ensuring unique assignment to a single worker.

In particular, up to 13 of the keypoints extracted in the pose estimation are used for the worker-PPE matching as follows: The central point of the two ears, the two eyes and the nose is used to identify the head and match it with the helmet. The center of the two shoulders and two hips is used for the vest and the protective suit. Finally, the wrists and the ankles are used for the gloves and the boots, respectively.

As mentioned above, in [6,7], pose estimation is used to define the regions of interest prior to PPE detection, whereas we use pose estimation to assign the PPE items to the right worker once they have been detected using distances from PPE to related body parts.

For cases where a PPE item does not directly align with any specific pose keypoint, it is assigned to the worker whose bounding box center is closest to the item, providing robust attribution even in complex scenes. While our approach ensures an important accuracy in PPE association, it does not always guarantee precise placement in some extreme cases (e.g., helmets may be near the head but not always perfectly positioned on it). Figs. 4 and 5 illustrate scenarios where workers are detected at medium and long distance from the camera, or when they are only partially visible. In Fig. 4, a worker and his helmet are identified at the bottom (only a portion of his head is visible, making PPE assignment challenging), yet the helmet is still correctly assigned. Meanwhile, the other two workers and their equipment are properly detected. In Fig. 5, workers and their attire are correctly detected and matched, including helmets, vests, and gloves, even in blurred sections of the image in a long-distance scene.

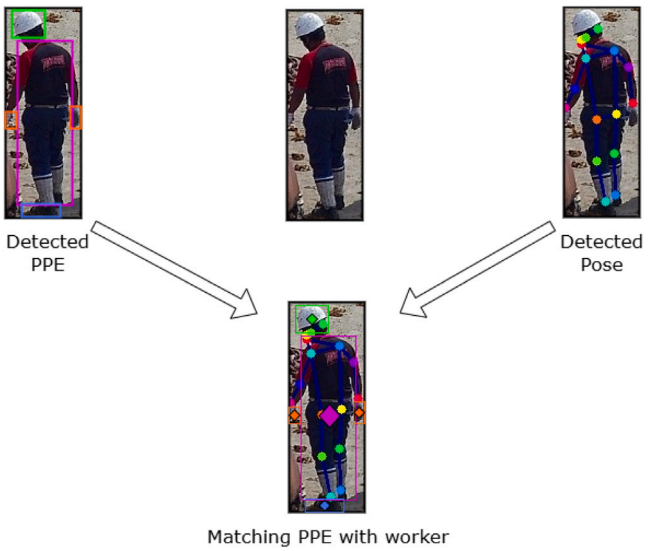

Fig. 3. Worker-PPE matching: in each cropped image (top middle), PPE is detected (top left) and pose is estimated (top right). The distance from the center of each PPE element (colored rhombi) to the corresponding body part (colored circles) is used for worker-PPE matching (bottom).

4. Experimental setup

In this section, we describe the experimental setup used to evaluate the proposed framework, detailing the characteristics of the datasets, as well as the metrics employed for performance assessment.

4.1. Collected dataset from PPE public repositories (training and validation dataset)

Based on the dataset published in [17],[7] we gathered, cleaned and combined multiple publicly available datasets with different classes in the field of PPE equipment.[8][9][10][11][12][13][14]

Afterward, 13,589 images were manually labeled to achieve a coherent distribution of PPE items and ensure consistent labeling across all categories. These images were split into training (10,207 samples) and validation (3380 samples) subsets, maintaining the original proportion of instances across each category. Additionally, 2198 worker samples from the Zamakona shipyard were added, with 1946 allocated to the training subset and 252 to the validation subset. Overall, the training subset included 12,153 individual worker images, while the validation subset contained 3632, each representing a single worker.

In Fig. 6, the number of instances per class in the collected dataset is shown. Although our dataset exhibits an unbalanced distribution of classes, we maintained this distribution in both the training and

[7] https://github.com/ZijianWang-ZW/PPE_detection.
[8] https://universe.roboflow.com/roboflow-100/construction-safety-gsnvb.
[9] https://universe.roboflow.com/nom/epi_detector.
[10] https://universe.roboflow.com/detection-kd8gd/fm4.
[11] https://universe.roboflow.com/temp04/ppe-8d778.
[12] https://universe.roboflow.com/ai-project-yolo/ppe-detection-q897z.
[13] https://universe.roboflow.com/universe-datasets/hard-hat-universe-0dy7t/dataset/6.
[14] https://universe.roboflow.com/helmet-tf/safety-helmet-4mhdt.

Fig. 4. Results for worker and PPE detection and matching in challenging scenarios: medium distance and minimally visible worker. Whole scene (left) and enlarged areas (right).

Fig. 5. Results for worker and PPE detection and matching in challenging scenarios: long distance. Whole scene (top) and enlarged areas (bottom).

validation processes. This approach was adopted to reflect the real-world scenario as accurately as possible and to ensure a consistent evaluation of the model's performance under realistic conditions (the imbalance arises from the different frequency in the images of some elements, such as safety boots and protective suits, when compared to helmets).

4.2. Dataset for workplace hazard prevention in a shipyard (Zamakona evaluation dataset)

In order to assess the effectiveness of the proposed framework, a benchmark dataset was created using high-resolution 1080p images captured at the Repnaval shipyard of Zamakona, located in the port of Las Palmas de Gran Canaria, Spain. The dataset consists of 227 images collected over several months using a Pan–Tilt–Zoom camera mounted on the rooftop of the shipyard's office building. With the intention of testing the model in especially difficult scenarios which highlight the improvements achieved, all images in the test dataset contain at least two overlapping bounding boxes, i.e., at least two workers are close to each other. In many cases, the number of nearby workers is higher. In addition, the images were captured from various perspectives and at different times of day, ensuring a diverse range of conditions in our dataset. This ensures that the difficulties associated with low light, shadows and poor visibility are also comprehensively considered.

Some samples are depicted in Fig. 7, showing the variety of lighting conditions, camera distances, and environmental scenarios (e.g., shadows, sunny or cloudy days) present in the collection. This diversity in angles and circumstances provides a comprehensive representation of the dataset and is crucial for conducting a robust evaluation of the proposed method across different real-world scenarios.

These images were manually labeled with the bounding boxes of the workers and visitors, along with the PPE classes stated before. In total, 1277 workers were labeled inside these images. In Fig. 8, the number of instances per class of the PPE associated to these workers is shown.

4.3. Performance evaluation metrics

The most commonly used performance metrics in this type of problems are precision (also known as positive predictive value) and recall

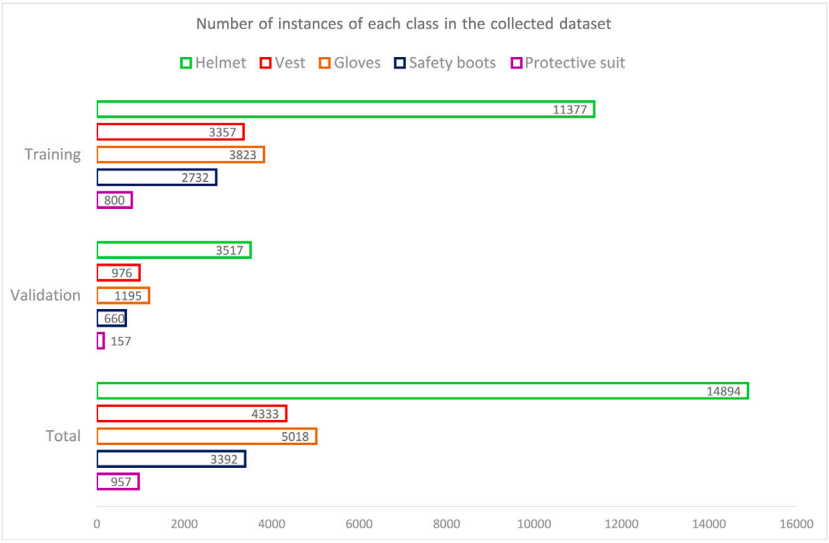

Fig. 6. Number of instances of each class in the collected dataset: training, validation, and total.

Fig. 7. Sample images from the Repnaval Zamakona dataset: eight diverse viewpoints that encompass various shipyard areas with different environmental conditions.

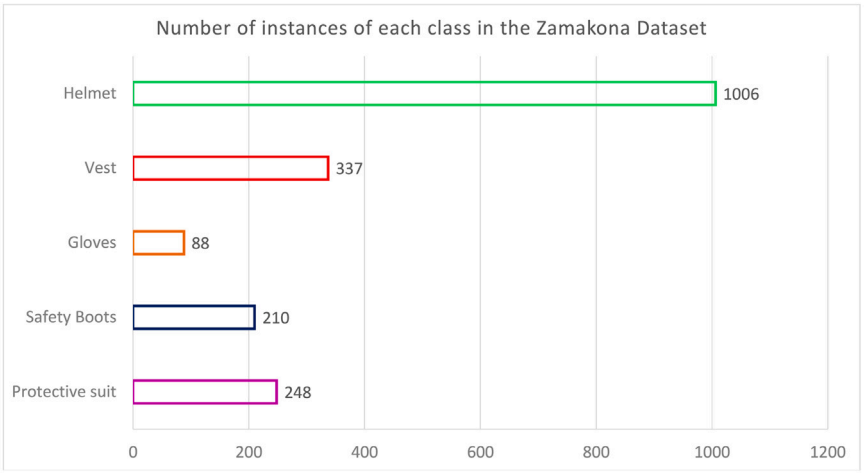

Fig. 8. Distribution of instances per class in the Zamakona evaluation dataset.

Table 2
Computation of precision and recall in multilabel classification. N: number of classes; M: number of instances. TP_i, FP_i, and FN_i: true positives, false positives, and false negatives for class i, respectively. $Precision_i$, $Recall_i$: precision and recall for class i. S_i: number of actual instances (support) for class i. $TP_{i,j}$, $FP_{i,j}$, and $FN_{i,j}$: true positives, false positives, and false negatives for class i in instance j.

	Precision	Recall
Micro	$\frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i + FP_i)}$	$\frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i + FN_i)}$
Macro	$\frac{1}{N}\sum_{i=1}^{N} Precision_i = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i}$	$\frac{1}{N}\sum_{i=1}^{N} Recall_i = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i}$
Weighted	$\frac{\sum_{i=1}^{N} Support_i \cdot Precision_i}{\sum_{i=1}^{N} Support_i}$	$\frac{\sum_{i=1}^{N} Support_i \cdot Recall_i}{\sum_{i=1}^{N} Support_i}$
Sample	$\frac{1}{M}\sum_{j=1}^{M} Precision_j = \frac{1}{M}\sum_{j=1}^{M}\frac{\sum_{i=1}^{N} TP_{i,j}}{\sum_{i=1}^{N}(TP_{i,j}+FP_{i,j})}$	$\frac{1}{M}\sum_{j=1}^{M} Recall_j = \frac{1}{M}\sum_{j=1}^{M}\frac{\sum_{i=1}^{N} TP_{i,j}}{\sum_{i=1}^{N}(TP_{i,j}+FN_{i,j})}$

Table 3
Computation of F1-score in multilabel classification. N: number of classes; M: number of instances. TP_i, FP_i, and FN_i: true positives, false positives, and false negatives for class i, respectively. $Precision_i$, $Recall_i$, and $F1_i$: precision and recall for class i. S_i: number of actual instances (support) for class i. $TP_{i,j}$, $FP_{i,j}$, and $FN_{i,j}$: true positives, false positives, and false negatives for class i in instance j.

	F1-score
Micro	$\frac{2 \cdot \text{Micro-Precision} \cdot \text{Micro-Recall}}{\text{Micro-Precision} + \text{Micro-Recall}}$
Macro	$\frac{1}{N}\sum_{i=1}^{N}\frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$
Weighted	$\frac{\sum_{i=1}^{N} \text{Support}_i \cdot \text{F1}_i}{\sum_{i=1}^{N} \text{Support}_i}$
Sample	$\frac{1}{M}\sum_{j=1}^{M}\frac{2 \cdot \text{Precision}_j \cdot \text{Recall}_j}{\text{Precision}_j + \text{Recall}_j}$

(also known as sensitivity). These metrics are derived from the counts of true positives (TP), false positives (FP), and false negatives (FN), and can be calculated as shown in Eq. (1). Note the $TP + FP$ represents the total number of retrieved instances, while $TP + FN$ is the total number of relevant instances. Therefore, precision reflects how specific the method is, while recall indicates how sensitive it is.

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

In classification tasks, it is common to combine both metrics in a harmonic mean called $F1$ score, as in Eq. (2).

$$\text{F1 score} = \frac{2TP}{2TP + FP + FN} \tag{2}$$

4.4. Multiclass combined evaluation

In single-label classification, a retrieved sample can be either correct or incorrect. However, in multi-label classification, a sample can also be partially correct (i.e., some elements of the PPE may be properly identified, while others are not). We could use a One-vs-Rest approach for each class to evaluate performance independently. However, to provide a more comprehensive assessment and address the unbalanced number of class instances, we have also evaluated performance using several averaging methods: micro, macro, weighted, and sample averages.

The expressions to calculate the average using these approaches are shown in Tables 2 (for precision and recall) and 3 (for F1-score). Micro-averaging aggregates all TP, FP and FN, across all classes to produce global values. This approach treats all retrieved elements (both true and false) and all non-retrieved actual elements equally, regardless of their class. Macro-averaging, in turn, calculates the metrics for each class individually and then computes the average of these metrics. This method ensures that each class contributes equally to the final results, regardless of the number of instances in the dataset or the number of retrieved elements. Weighted averaging considers the number of instances for each class, known as the class's support (S). The metric for each class is weighted by its support, thus meaning that classes with more occurrences have a higher influence on the final metric. Finally, sample averaging computes the metrics for each sample across the different classes, and then averages the values across all samples. Consequently, this approach gives equal contribution to all samples, regardless of the number of classes present in them.

These metrics should be used in conjunction, especially in unbalanced datasets like ours. While micro-averaging may downplay the impact of less-represented classes, macro-averaging can exaggerate poor performance in those classes. Weighted averaging aims to strike a balance by considering the number of actual instances, but does not fully address class imbalance. Finally, sample averaging provides insight into how the model performs on individual samples, regardless of the number of classes which are contained.

5. Experimental results

In this section, the main results are presented and discussed. First, the training method and its configuration are explained. Afterwards, the method is compared with state-of-art research, testing different combinations within the general framework. Finally, an ablation study

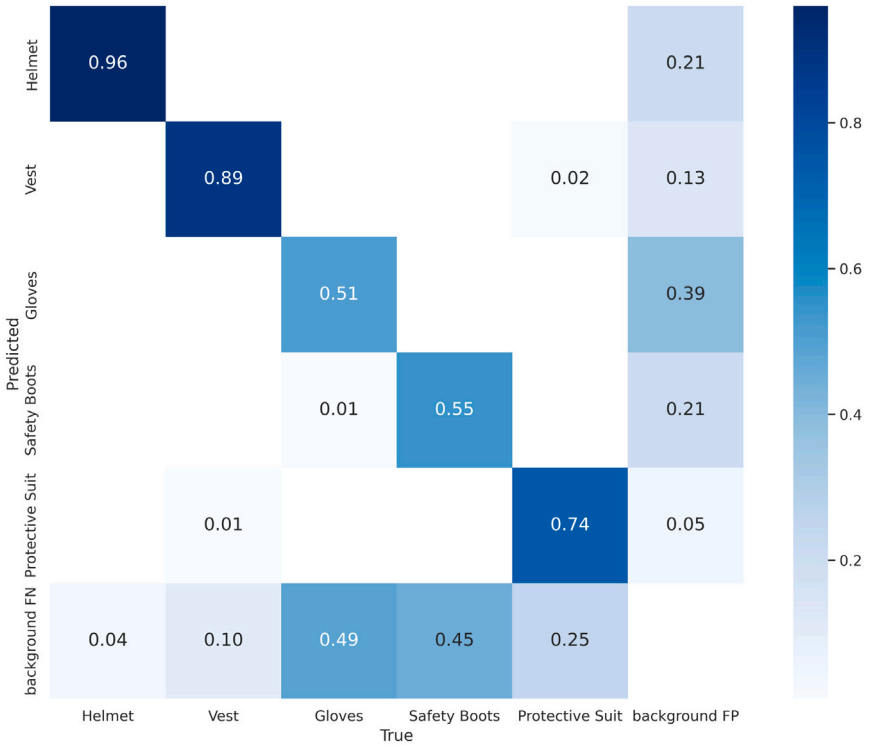

Fig. 9. Confusion matrix illustrating the classification performance across the PPE classes.

is conducted to show the improvements provided by the different stages of the proposed framework.

The framework we have presented consists of four main stages, namely worker detection, PPE detection, pose estimation, and worker-PPE matching. For worker detection, we work with $192x192$ images, and the configuration can be changed to use a different interpolation method or to increase/decrease the margins of the bounding boxes. However, the probability of overlapping increases if margins are large, so that they must be cautiously adjusted. For both worker and PPE detection, the threshold for the confidence of the detection is set to 0.5. If this threshold is increased, fewer elements are identified and, if it is reduced, the identified ones are less reliable. Therefore, a balance must be achieved. Regarding PPE identification, we work with five categories (helmet, vest, gloves, boots and protective suits). Using too many categories would drastically increase the possible configurations of present/absent PPE items, so that the most significant ones must be chosen. In this stage, the overlapping for duplicate removal must be controlled, and this requires a threshold for the IoU (θ_o, which has been set to 0.65 in our experiments). Adjusting this value allows balancing the duplicate removal with the possibility of removing an actual original instance. Pose estimation identifies 17 body parts, although, in our case, 13 are considered, since they are the relevant ones for the PPE items we work with. In fact, the worker-PPE matching deals with 6 body parts (head for the helmet, chest for the vest and protective suit, hands for the gloves, and feet for the boots, each one estimated from a set of pose keypoints, as described in Section 3.4) and calculates the distance from PPE items to the corresponding body part.

The experiments were carried out in a computer with the following features: Intel® Xeon® Gold 6230 Processor @2.10 GHz, GPU RTX 3060 cores 3584 @1.78 GHz 12 GB GDDR6, RAM 62 GB DDR4 @2400MH.

5.1. Results for the different classes in the PPE detector

To implement the proposed framework, we trained a Yolov7, variant e6, for the PPE detection step, using the dataset described in Section 4.1 for training and validation. Regarding the results in the validation set, Fig. 9 illustrates the confusion matrix for the five PPE classes, indicating the proportion of instances of a class which are assigned each label or interpreted as background. As observed, helmets and vests are well identified in most cases with correct predictions of 96% and 89%, respectively. However, approximately a quarter of the protective suits and almost half of the gloves and safety boots are not identified. This was to be expected, due to the challenges posed by their sizes, colors and positions.

Fig. 10 presents the precision–recall curve, plotting the metrics for various score cutoffs, and illustrating the trade-off between precision and recall. As observed, the area under the curve is quite large for both helmets and vests (0.970 and 0.888, respectively). This indicates that a high precision can be achieved for these classes without significantly compromising recall, and vice versa. For protective suits, increasing recall beyond 60% would require a moderate reduction in precision. Finally, gloves and safety boots pose the greatest challenge, with only modest levels of precision and recall achievable.

Table 4 provides detailed values of precision and recall, as well as the average precision (AP) for each of the 5 classes and for the combination of all of them. In this table, AP_{50} indicates the average precision when the threshold IoU (Intersection over Union between the predicted and the actual instances) is set to 50%, so that a result is considered a true positive when the predicted and the actual instances overlap in at least a half. On the other hand, the mean average precision (mAP, also denoted as AP@[.50;.05:.95], or simply AP) reflects the mean of the average precision across IoU thresholds ranging from 50% to 95% in 5% increments.

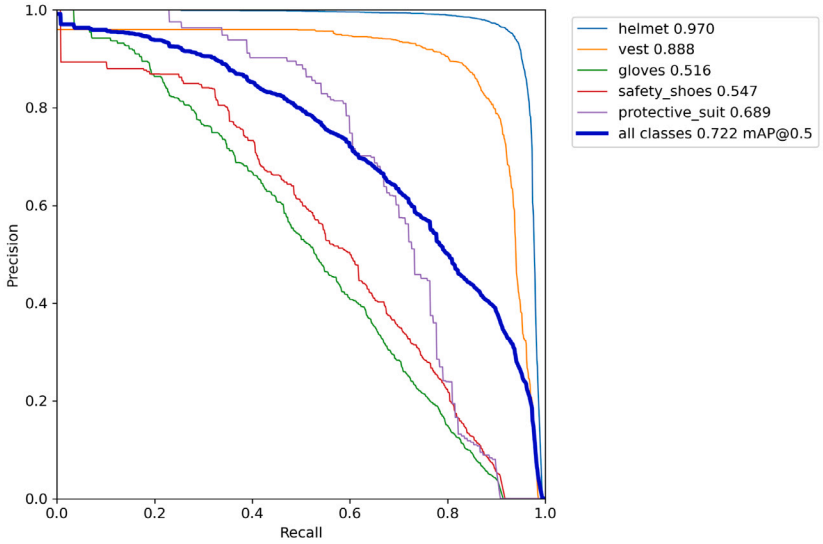

Fig. 10. Precision–Recall curve obtained from the evaluation in the Zamakona dataset.

Table 4
Results for the validation dataset (3632 images): precision (P), recall (R), average precision for threshold IoU 50% (AP_{50}), and mean of average precision for range of thresholds 50%–95% (AP).

Class	Labels	P	R	AP_{50}	AP
All	6503	0.735	0.684	0.722	0.445
Helmet	3515	0.928	0.946	0.970	0.697
Vest	976	0.829	0.874	0.888	0.552
Gloves	1195	0.614	0.449	0.516	0.264
Safety boots	660	0.643	0.479	0.547	0.271
Protective suit	157	0.662	0.669	0.689	0.441

Table 5
Results for worker multilabel assignation on our benchmark dataset. Individual metrics for each PPE class and aggregated metrics using different averaging strategies.

Class	Metrics			
	Precision	Recall	F1-score	support
Gloves	39.32	52.27	44.87	88
Helmet	93.16	73.06	81.89	1006
Protective suit	63.64	62.10	62.86	248
Safety boots	59.32	83.33	69.31	210
Vest	82.27	68.84	74.96	337
Accumulated average:				
Micro avg.	77.80	71.04	74.27	1889
Macro avg.	67.54	67.92	66.78	1889
Weighted avg.	81.07	71.04	75.03	1889
Sample avg.	53.66	57.12	53.71	1889

5.2. Evaluation on the Zamakona dataset

Table 5 contains the metrics resulting from the evaluation of our method in terms of attribute assignation as a multi-label task. When considering the different elements of the PPE individually, the highest precision is achieved for helmets (93.16%) and vests (82.27%). Although not as high, precision is moderately good for protective suits and safety boots, while it is lower for gloves. Concerning recall, it is satisfactory for safety boots and helmets (over 70%), and moderate for vests, protective suits and gloves. Note that recall is affected by the number of non-detected workers, as each element in their equipment will count as a false negative for the corresponding class. This results in higher F1-scores for helmets and vests, moderate values for safety boots and protective suits, and lower figures for gloves.

When combining all instances into a single metric, the fact that helmets are the most frequent class, while gloves are the least common one, penalizes the macro average and the sample average. On the contrary, the micro and the weighted averages are improved (approximately 80% precision, 70% recall, and 75% F1-score).

5.3. Quantitative comparison with state-of-the-art methods

In this section, we present a quantitative comparison of the proposed framework with the approaches introduced by Nath et al. [8] and our own implementation from the one presented in Xiong et al. [7]. To maintain clarity, we will refer to these methods simply as Nath and Xiong throughout the rest of the paper.

Nath proposed three strategies trained on the Pictor dataset, labeled A1, A2, and A3. The last two, Nath-A2 and Nath-A3, share similarities with our framework and are thus suitable for comparison. Nath-A2 employs a single-stage detection model designed to simultaneously identify workers and classify them based on their PPE compliance. Each detected worker is categorized into one of four groups: without PPE, with a helmet, with a vest, or with both a helmet and a vest. Nath-A3, on the other hand, follows a two-stage approach. First, a detector identifies workers, and then a multi-class classifier assigns them to the same PPE categories as in Nath-A2.

Additionally, Xiong introduces a PPE compliance monitoring system that, like our approach, uses pose estimation to guide PPE detections. However, PPE items are searched around the corresponding body part. Thus, Xiong's approach serves as a useful benchmark for comparing strategies in PPE detection and assignment.

Tables 6 and 7 present the metrics comparing the worker detector that we use (InternImage), against the YOLOv3-based approaches proposed by Nath et al. [8] (A2 and A3). In their dataset (Pictor), InternImage outperforms both YOLOv3-based models across all metrics, achieving a Precision of 92.80%, Recall of 83.18%, and an F1-score of 87.73%. In contrast, the best-performing YOLOv3-based model (Nath A3) reaches an F1-score of 82.83%, with lower precision and recall values. This improvement highlights the enhanced detection capabilities of InternImage, particularly in accurately identifying workers while maintaining a high recall rate. The difference is even more pronounced in the Zamakona dataset, where the YOLOv3-based models struggle significantly. Nath A2 and A3 exhibit low recall values (35.71% and 40.09%), leading to poor F1-scores (45.83% and 49.47%), indicating frequent missed detections. In contrast, InternImage achieves a much higher recall (76.74%) and an F1-score of 85.33%, demonstrating its robustness across different datasets.

Table 6
Worker detection metrics in the Pictor dataset.

Model	Precision	Recall	F1-score
Nath A2 (YOLOv3)	87.70	77.19	82.11
Nath A3 (YOLOv3)	86.47	79.49	82.83
Ours (InternImage)	92.80	83.18	87.73

Table 7
Worker detection metrics in the Zamakona dataset.

Model	Precision	Recall	F1-score
Nath A2 (YOLOv3)	63.96	35.71	45.83
Nath A3 (YOLOv3)	64.56	40.09	49.47
Ours (InternImage)	96.08	76.74	85.33

This difference highlights the advancements in object detection models, training techniques, and datasets over the years, leading to significantly better results for modern top-performing models compared to older architectures like YOLOv3. The performance of the worker detector is crucial for the overall efficacy of both PPE detection and assignment tasks.

Regarding PPE detection and assignment, we also conducted a comparison against Xiong and Nath A2 and A3, evaluating their different variants on both the Pictor dataset and the Zamakona dataset. Although our framework is capable of detecting five different PPE categories, the other approaches are limited to detecting only helmets and vests. Therefore, the extracted metrics for this comparison focus exclusively on these two categories.

As previously mentioned, worker detection plays a crucial role in the effectiveness of both methods. The results indicate that the original version of Nath's approach performs worse than our method on its own dataset (Pictor dataset), despite our method not being trained on it, and performs significantly worse on the Zamakona dataset. To ensure a fairer comparison of PPE detection and worker-PPE matching, we also tested their method after replacing YOLOv3 with InternImage as the worker detector. This substitution resulted in a significant improvement in performance of their original proposal on both datasets.

Table 8 presents the combined performance metrics when applying the state-of-the-art models and our proposed framework to our dataset and the Pictor dataset. As observed, our method outperforms both Nath and Xiong by a significant margin across all metrics in the Zamakona dataset. The precision, recall, and F1-score of our method exceed the best results obtained with the others approaches in all averaging options, with some improvements surpassing 20%.

On the other hand, when tested on the Pictor dataset, our method achieves comparable results. The best results for Nath's models on their own dataset exceed our method by less than 5% in micro, weighted, and sample averages. However, while their method achieves a higher macro average precision, this comes at the expense of a lower recall, with our method surpassing theirs by more than 5% in recall. Additionally, our method attains slightly higher precision in the weighted average. Overall, our proposal consistently outperforms the Xiong approach, particularly by normally achieving higher recall rates, which in turn leads to a better F1 score. Despite the Pictor dataset containing images with closer distances and being trained for only two PPE categories, our general-purpose framework delivers competitive results.

Finally, in Table 9 we compare the mean inference times of the PPE detection models on the Zamakona evaluation dataset. For a fair comparison, we have used the models trained on InternImage for Nath A3. Nath-A2 achieves the fastest time at 0.124 s per image due to its simpler architecture, which integrates worker detection and PPE classification into a single step, reducing computational demands, but sacrificing feature extraction capabilities and providing poor results. Nath-A3 models strongly enhance metrics compared to Nath-A2, thanks to the superior feature representation provided by their deeper architectures, though this comes at the cost of longer inference times. The Xception variant is slightly faster than ResNet, benefiting from depthwise separable convolutions that maintain feature quality while reducing computational load, thereby improving performance in complex scenarios.

Xiong's method has a longer processing time (0.964 s) compared to the other approaches due to its multi-stage pipeline, which includes worker detection, pose estimation, and separate classifiers for PPE detection. While this strategy enhances PPE localization accuracy, it also increases computational overhead. The sequential processing of worker detection followed by region-of-interest extraction based on estimated body parts before PPE classification contributes to this time. This reliance on pose estimation adds complexity and affects overall detection speed, making it slower than the other evaluated methods.

In contrast, our framework operates at 0.878 s per image while incorporating additional steps such as collision detection and duplicate removal to refine PPE assignment. Unlike Xiong, our method does not rely on pose estimation to define PPE search areas; instead, it directly associates detected PPE with the nearest worker while leveraging pose information for final matching. This process reduces computational overhead. Additionally, our framework balances efficiency by integrating pose estimation in a way that enhances detection without introducing excessive delays.

These results demonstrate a trade-off between inference speed and detection accuracy. While Nath-A2 offers faster processing, its simplified architecture compromises precision. Conversely, deeper models such as Nath-A3 and our framework provide significantly improved detection and matching performance. Our method, in particular, demonstrates important accuracy in real-world construction site environments, effectively covering broad areas while maintaining competitive inference speed suitable for practical deployment.

5.4. Qualitative comparison with state-of-the-art methods

In this section, we present a qualitative comparison of the proposed method against Xiong [7] and the enhanced implementation of Nath [8]. We aim to evaluate the strengths and limitations of each PPE detection framework across different conditions, particularly focusing on occlusions, low illumination, and non-compliance scenarios. This comparative analysis provides insights into the effectiveness of each method and identifies areas for further improvement in PPE detection.

Besides, in order to further validate the generalizability of our framework, we extend our evaluation beyond the Zamakona dataset by testing on additional real-world construction site images. These images include new, previously unseen backgrounds and environmental conditions, ensuring that our model is not biased toward a specific dataset. The evaluation includes images from the Nelson Mandela dock in the Port of Las Palmas de Gran Canaria, where dock extension work is underway, as well as real construction sites in Gran Canaria and publicly available images from Shutterstock.[15] These datasets were chosen to introduce greater variability in construction environments, such as different lighting conditions, worker densities, and PPE compliance levels.

In each figure, we facilitate visual analysis by using a color-coded bounding box system to differentiate PPE compliance levels. A white bounding box indicates a detected worker without any PPE. A green bounding box indicates a worker wearing a helmet, a red bounding box represents a worker wearing a vest, and a yellow bounding box denotes a worker equipped with both a helmet and a vest. This visualization scheme provides an intuitive representation of PPE compliance status across various scenarios. Notice that our framework is capable of detecting a broader range of PPE categories compared to the other methods. However, for this evaluation, we focus on the main PPE classes detected by all three approaches to ensure a fair and consistent comparison.

[15] https://www.shutterstock.com/.

Table 8
Comparison of our method against state-of-the-art models on the Zamakona and Pictor datasets. Bold values indicate the best in each column; underlined values mark the best among existing methods. The last row in each section shows the difference between our results and the best alternative.

Detector	Model	Micro			Macro			Weighted			Sample		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
YOLO	Nath-a2	57.28	26.65	36.38	48.27	18.77	24.43	53.16	26.65	33.86	22.52	19.13	20.18
	Nath-a3-vgg	57.51	29.33	38.85	62.68	25.69	35.27	59.20	29.33	38.35	22.62	19.64	20.58
	Nath-a3-resnet	57.55	32.91	41.87	56.07	29.96	38.84	57.28	32.91	41.65	23.74	21.91	22.18
	Nath-a3-xception	61.43	26.20	36.74	61.31	25.09	35.55	61.41	26.20	36.70	18.00	16.84	17.05
II	Nath-a3-vgg	83.41	54.79	66.13	73.17	43.50	53.00	80.10	54.79	64.04	51.90	46.77	48.01
	Nath-a3-resnet	76.08	56.92	65.12	67.07	49.85	57.18	75.72	56.92	64.98	48.56	47.00	46.47
	Nath-a3-xception	80.62	44.60	57.43	75.84	41.71	53.82	80.51	44.60	57.40	38.15	35.65	35.99
	Xiong et al. [7]	83.65	42.29	56.18	85.07	41.55	55.78	83.77	42.29	56.17	34.98	32.56	33.26
II	Ours	**90.29**	**72.00**	**80.12**	**87.71**	**70.95**	**78.43**	**90.42**	**72.00**	**80.15**	**57.25**	**57.82**	**57.10**
Difference		+6.64	+15.08	+13.99	+2.64	+21.10	+21.25	+6.65	+15.08	+15.17	+5.35	+10.82	+9.09
YOLO	Nath-a2	76.81	72.85	74.77	69.86	72.15	70.83	76.88	72.85	74.79	42.93	42.93	42.82
	Nath-a3-vgg	76.70	73.53	75.08	66.48	72.51	68.97	76.88	73.53	75.13	42.62	42.72	42.62
	Nath-a3-resnet	72.63	73.88	73.25	54.25	79.65	61.18	74.17	73.88	73.74	42.31	42.82	42.48
	Nath-a3-xception	74.25	68.38	71.19	51.19	62.90	53.77	76.12	68.38	71.84	39.44	39.85	39.54
II	Nath-a3-vgg	**83.15**	**76.29**	**79.57**	65.11	73.92	**67.96**	83.82	**76.29**	**79.77**	**46.86**	**46.75**	**46.75**
	Nath-a3-resnet	74.83	75.60	75.21	53.48	80.53	59.85	77.19	75.60	76.02	45.67	46.32	45.82
	Nath-a3-xception	80.75	73.54	76.98	59.52	65.54	61.26	81.56	73.54	77.26	45.35	45.13	45.17
	Xiong et al. [7]	78.17	61.51	68.85	67.32	66.35	65.77	78.52	61.51	68.89	33.46	33.46	33.40
II	Ours	78.31	73.20	75.67	53.57	**86.27**	56.54	**84.97**	73.20	77.85	42.86	44.59	43.43
Difference		−4.84	−3.09	−3.90	−13.75	+5.74	−11.42	+1.15	−3.09	−1.92	−4.00	−2.17	−3.33

Fig. 11. Comparison of the results obtained when workers are close to each other and their bounding boxes overlap: our method (left), Nath's (middle), and Xiong's (right). For each method: result for the whole scene (top), and enlargement of the selected area (bottom).
Worker with no PPE only helmet only vest helmet and vest.

Table 9
Comparison of mean inference speed between Nath and Xiong models and our proposed framework.

Model	Seconds
Nath-a2	0.124
Nath-a3-vgg	0.706
Nath-a3-resnet	0.852
Nath-a3-xception	0.771
Xiong	0.964
Ours	0.878

Inter-worker occlusions: In Fig. 11, a particularly difficult example of overlapping workers is illustrated. As shown in the enlarged areas, our method detects all four workers, including one who is largely occluded by two other workers. Furthermore, their helmets are properly identified, as the three vests which are worn by the workers (one of them is not wearing a vest). Nath's method detects all workers, but misses one of the helmets and wrongly assigns a vest to a worker who is not wearing it (in fact, a vest is duplicated and assigned to two workers). Finally, Xiong's is not capable of identifying the helmet and vest of one of the workers and mistakenly assigns a vest to another one. This highlights the importance of the collision detection and duplicate removal stage which we have included in our approach, since this phase avoids assigning a PPE item to a worker who is not wearing it when they are close to another worker who has it. Furthermore, the use of the pose estimation helps matching the item with the most probable owner. In Fig. 12, another example with overlapping bounding boxes is shown in the bottom subimages. Xiong's approach identifies two workers, but misses the worker in the middle, who is occluded by the others, and one helmet. Nath's method identifies the three workers, but concludes that only one of them is wearing a hat, which is wrong. However, our method detects the three workers and assigns a hat to one of them, a vest to the second one, and hat and vest to the third one. In the enlarged area shown on top, we can see how our method is the only one capable to detect the helmet of the worker in the center while the other methods miss some helmets (Nath) or even workers (Xiong).

Densely occupied areas: In Fig. 13, 18 workers are visible. Our method identifies all of them, although fails to detect some of the helmets (top subimages). Nath's also detects all workers and performs

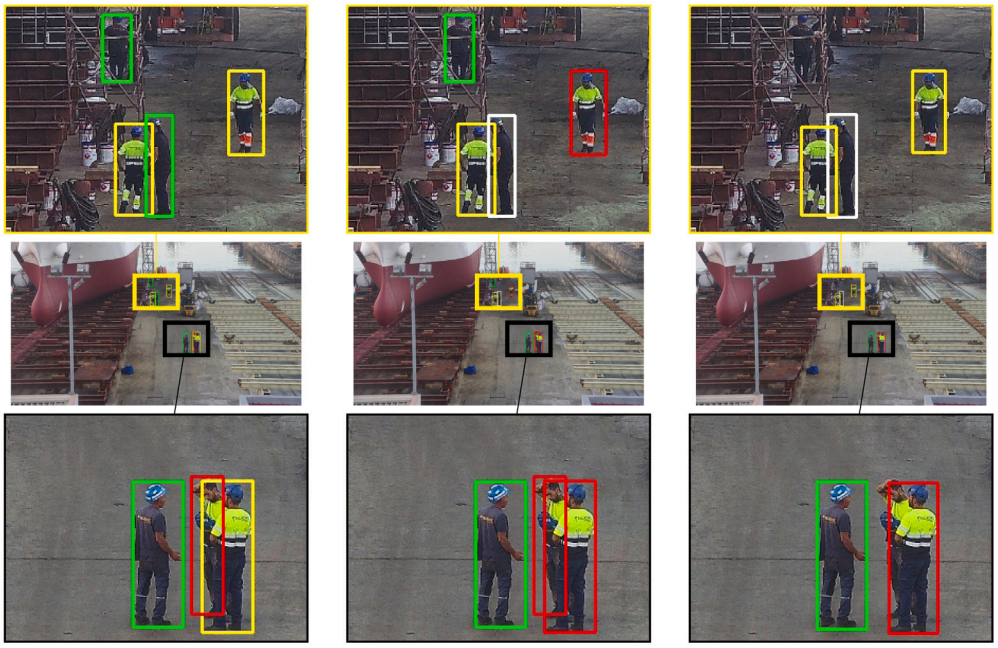

Fig. 12. Comparison of the results obtained when workers are close to each other and their bounding boxes overlap: our method (left), Nath's (middle), and Xiong's (right). For each method: result for the whole scene (middle row), and enlargement of the selected areas (top and bottom).
Worker with no PPE only helmet only vest helmet and vest.

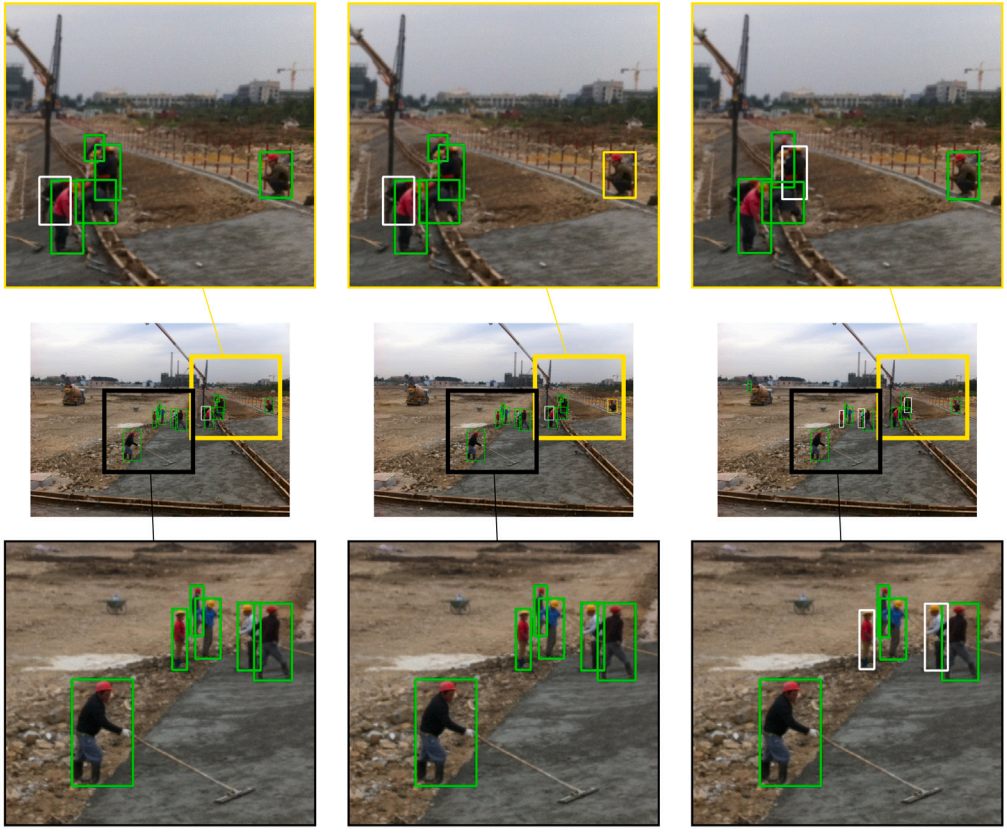

Fig. 13. Comparison of the results obtained when numerous workers are visible in the scene and some areas are densely occupied: our method (left), Nath's (middle), and Xiong's (right). For each method: result for the whole scene (middle row), and enlargement of the selected areas (top and bottom).
Worker with no PPE only helmet only vest helmet and vest.

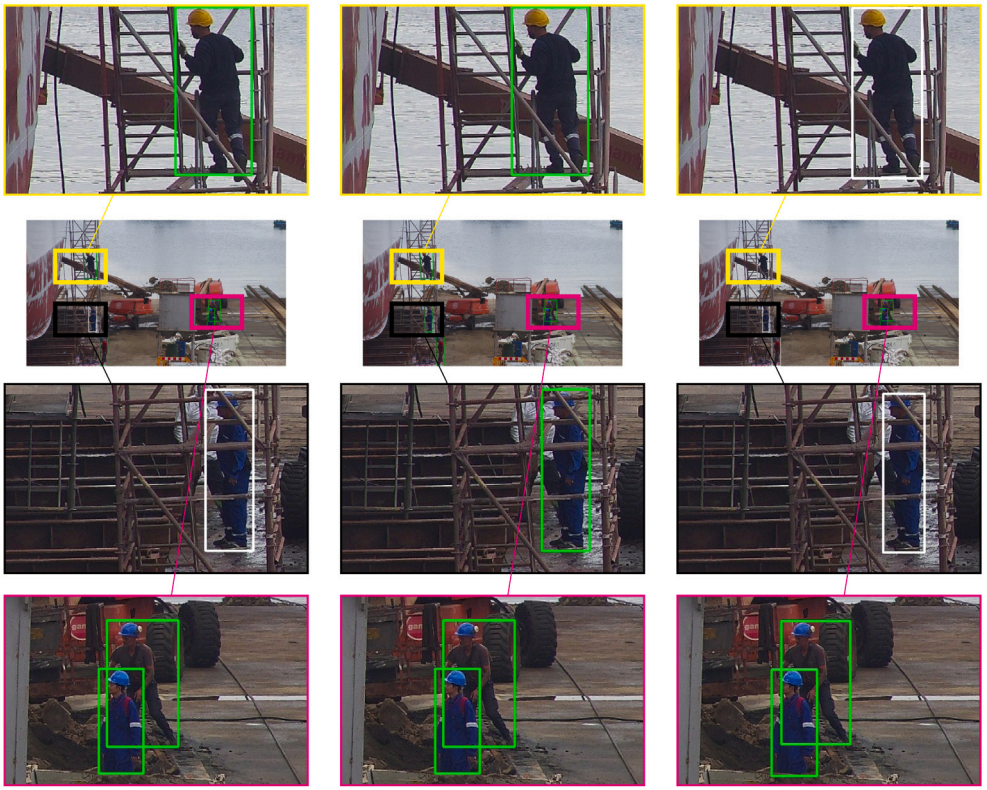

Fig. 14. Comparison of the results obtained when workers are occluded by scaffolding or other workers: our method (left), Nath's (middle), and Xiong's (right). Whole scene (2nd row) and enlargements (1st, 3rd and 4th rows).
Worker with no PPE only helmet only vest helmet and vest.

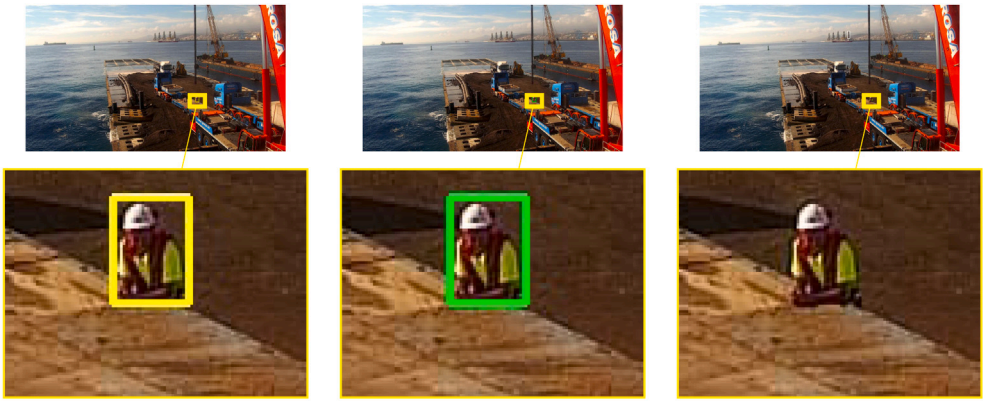

Fig. 15. Comparison of the results obtained when workers are only partially visible: our method (left), Nath's (middle), and Xiong's (right). For each method: result for the whole scene (top), and enlargement of the selected area (bottom).
No PPE helmet only vest only helmet and vest.

better in detecting the helmets, but wrongly assigns a vest to one of the workers (top subimage on the right). Finally, Xiong's misses some of the workers.

Occluding objects: In Fig. 14, occlusions between workers and scaffolding play an important role. In this case, our method detects all workers and only one helmet is missing (worker occluded by the scaffolding on the third row). Nath's approach detects all workers and their helmets, while Xiong's method misses two helmets. Although our method detects other items, such as protective suits, they are not shown here to compare the results with the other methods in this tests, which focus on helmets and vests. In Fig. 15, a worker beyond a truck is not completely visible. Our method detects him as well as his helmet and vest. Nath's only detects the worker and his helmet, and Xiong's does not detect the worker.

Challenging lighting conditions: Fig. 16 shows a comparison of the results in shadowed areas with occlusions. As observed, all three methods miss one of the workers in the top images, since only their helmet is visible. Our method identifies five workers (three on the top and two on the bottom subimages), and assigns a helmet to four of them and vest to one of them (reflective elements are considered in the category of vest). Nath's method also detects five workers, assigns a helmet to those in the shadowed area and no vest is identified. Finally, Xiong's method only detects two workers in the shadowed area. In Fig. 17, low-light makes it difficult to identify all elements. Nath's method fails to detect one of the helmets (worker on the left) and Xiong's misses

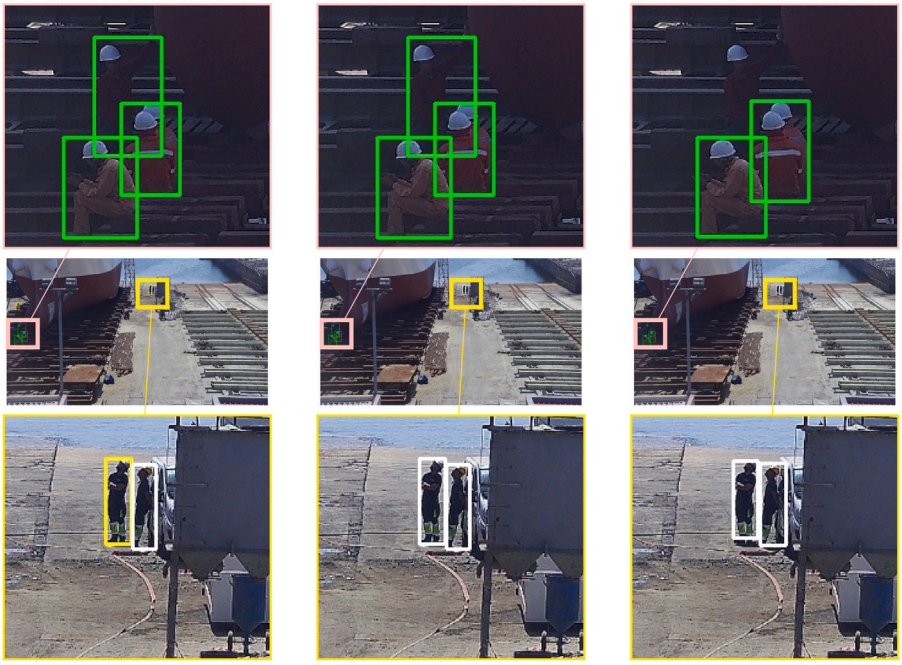

Fig. 16. Comparison of the results obtained in shadowed areas and with occluded workers: our method (left), Nath's (middle), and Xiong's (right). For each method: result for the whole scene (middle row), and enlargement of selected areas (top and bottom).
Worker with no PPE only helmet only vest helmet and vest.

Fig. 17. Comparison of the results obtained under low-light conditions and overlapping workers: our method (left), Nath's (middle), and Xiong's (right). For each method: result for the whole scene (top), and enlargement of the selected area (bottom).
Worker with no PPE only helmet only vest helmet and vest.

Fig. 18. Examples illustrating the methods' performance in untidy scenarios: our method (left), Nath's (middle), and Xiong's (right).
Worker with no PPE only helmet only vest helmet and vest.

a vest (right) and a helmet (middle), while our method properly detects all workers and PPE items.

Untidy scenarios: Fig. 18 includes and example of untidy and not well-organized scenario. All workers are properly detected by all three methods. However, some differences are noticeable in the detection of the PPE. Xiong's method wrongly assigns a helmet to one of the workers (top back), while it is not totally clear whether the closest worker wears a helmet or not. **Non-compliance scenes:** Fig. 19 (bottom) illustrates

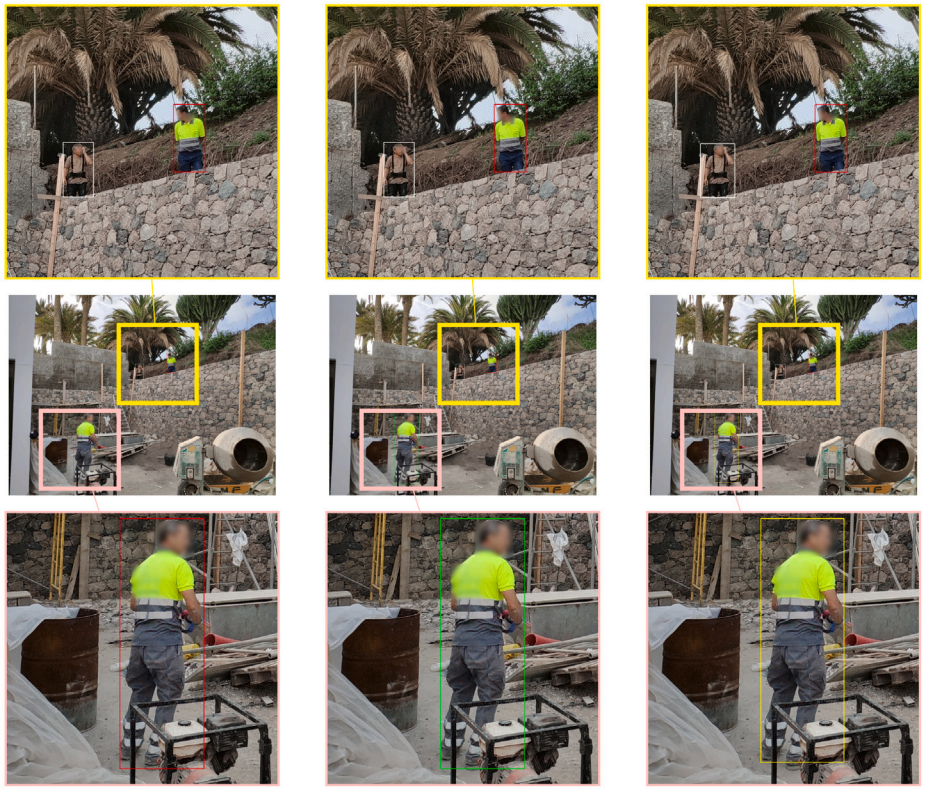

Fig. 19. Comparison of the results when workers do not comply with safety regulations: our method (left), Nath's (middle), and Xiong's (right). For each method: result for the whole scene (middle), and enlargement of selected areas (top and bottom).
Worker with no PPE only helmet only vest helmet and vest.

how the methods can err on the side of excess. Although the worker on the right wears a vest but no helmet, Nath's and Xiong's methods assign him a helmet. This does not happen with our proposal, which correctly indicates that the worker is not wearing a hat.

Partial failures in extremely difficult scenes: In Fig. 20 we show an example in which all methods partially fail. One of the workers is not detected by our method (bottom row), as happens with Nath's method, while Xiong's detects him, but identifies a bin as another worker. In the enlarged area on top, Xiong's cannot detect one of the workers, while Nath's and ours miss some of the equipment. Finally, in the third row, one of the workers is not detected by any method and all of them fail to detect some of the attire.

5.5. Evaluation of components for the global framework

In order to evaluate how the global result is affected when one of the components is substituted, we have tested the method on both datasets using different combinations of detectors and pose estimators. Note that Yolov7 is kept as PPE detector in every combination.

The results presented in Table 10 demonstrate that InternImage consistently achieves the highest performance across all evaluation metrics, irrespective of the pose estimation model employed. This trend holds true for both the Zamakona and the Pictor dataset evaluated, underscoring InternImage's robustness and effectiveness as a detection architecture. Compared to other detectors such as DDETR [37], Dino [38], and DDQ [39], InternImage achieves superior precision, recall, and F1 scores in most of the Micro, Macro, Weighted, and Sample-based evaluations.

Although the choice of pose estimation model introduces some variation in the results, these differences are relatively minor when contrasted with the impact of the detector. For instance, ViTPose exhibits slightly higher recall values compared to Hrnet [40] and SimCC [41], but the overall performance differences are slight. This suggests that, although pose estimation models do play a role in the final outcomes, the influence of their choice is significantly less substantial compared to the detector. A notable observation is that InternImage maintains its performance advantage even when paired with different pose estimation models, highlighting its consistent efficacy.

In conclusion, the findings from both the Zamakona dataset and the Pictor dataset reinforce the critical role of the detection architecture in determining overall performance. InternImage emerges as the key driver of accuracy improvement and, while pose estimation models contribute to the final result, substituting the pose estimator with a different model (among the tested estimators) does not affect the result as much as a change in the detector. These results emphasize the necessity of selecting a robust detector to achieve superior outcomes across various evaluation metrics.

5.6. Ablation study

In this section, we present a comprehensive ablation study to analyze the framework and underscore the importance of the pose estimation and duplicate removal modules. Table 11 provides a detailed summary of the results from our ablation study in terms of precision (P), recall (R), and F1-score (F1) across micro, macro, weighted, and sample averaging categories. The impact on performance of excluding the collision detection and duplicate removal (CDDR) and the pose estimation (Pose) components is illustrated, either individually or in combination.

With both CDDR and pose estimation enabled, the model achieves the highest performance across key metrics. The micro, macro, and weighted F1-scores reach 69.79, 63.29, and 70.61, respectively, outperforming all other tested configurations. Additionally, this configuration

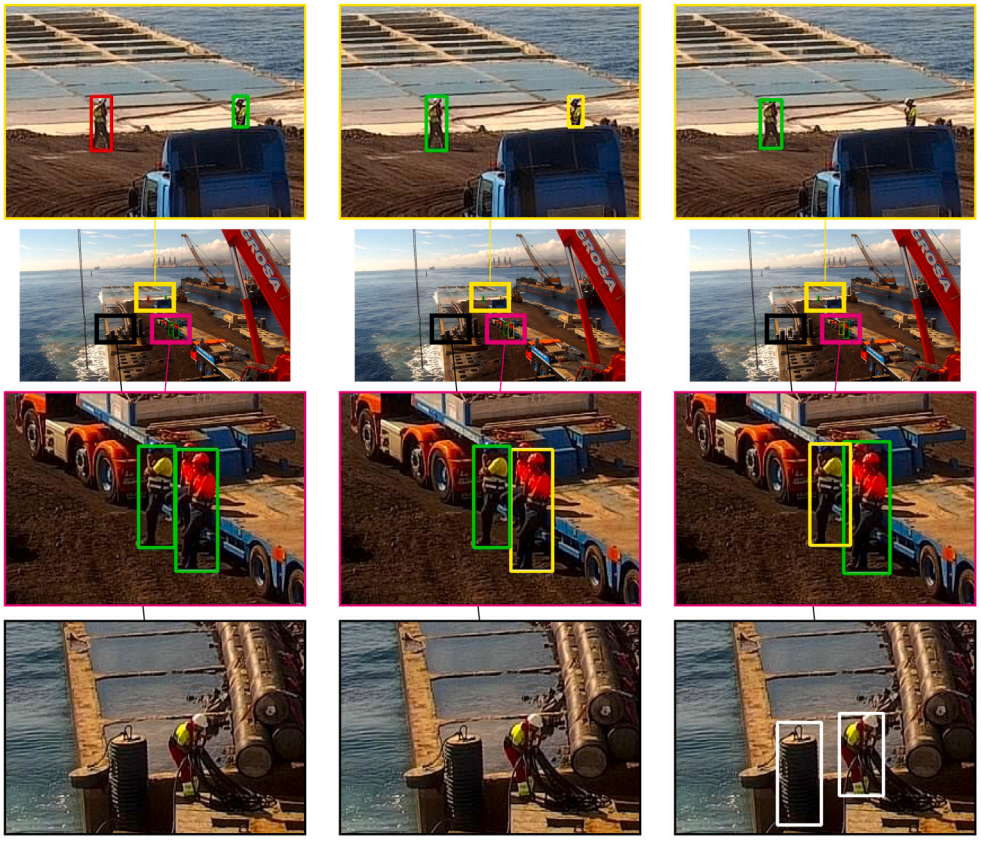

Fig. 20. Results in extremely difficult scenes with low luminance, partially occluded workers, long distances: our method (left), Nath's (middle), and Xiong's (right). For each method: result for the whole scene (2nd row), and enlargement of selected areas (1st, 3rd and 4th rows).
Worker with no PPE only helmet only vest helmet and vest.

Table 10
Evaluation of the performance of different combinations of detectors and pose estimators on Zamakona dataset (top section) and Pictor dataset (bottom section).

Person	Pose	Micro			Macro			Weighted			Sample		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
DDETR [37]	ViTPose [36]	88.76	58.22	70.32	85.79	57.71	68.95	89.05	58.22	70.37	46.01	46.47	45.85
DINO [38]	ViTPose [36]	90.13	56.44	69.41	87.47	55.82	68.12	90.32	56.44	69.44	45.20	45.74	45.16
DDQ [39]	ViTPose [36]	90.13	66.64	76.62	**87.96**	66.18	75.51	90.26	66.64	76.65	53.67	53.79	53.33
II [42]	Hrnet [40]	89.89	70.88	79.26	86.87	69.61	77.26	90.08	70.88	79.31	56.37	57.02	56.23
II [42]	SimCC [41]	90.05	70.81	79.28	87.10	69.66	77.38	90.24	70.81	79.33	56.37	56.98	56.20
II [42]	ViTPose [36]	**90.29**	**72.00**	**80.12**	87.71	**70.95**	**78.43**	**90.42**	**72.00**	**80.15**	**57.25**	**57.82**	**57.10**
DDETR [37]	ViTPose [36]	79.38	62.19	69.74	**55.60**	80.63	55.01	86.32	62.19	71.37	36.84	38.15	37.28
DINO [38]	ViTPose [36]	**80.64**	60.13	68.89	53.99	65.64	50.57	**87.76**	60.13	70.71	36.96	38.54	37.49
DDQ [39]	ViTPose [36]	78.62	70.79	74.50	54.55	85.03	**57.48**	84.27	70.79	76.18	42.10	43.64	42.61
II [42]	Hrnet [40]	78.30	73.19	75.66	53.57	86.26	56.53	84.96	73.19	**77.85**	**42.95**	**44.68**	**43.52**
II [42]	SimCC [41]	78.02	73.19	75.53	53.39	86.26	56.46	84.61	73.19	77.70	42.86	44.58	43.42
II [42]	ViTPose [36]	78.31	**73.20**	**75.67**	53.57	**86.27**	56.54	84.97	**73.20**	**77.85**	42.86	44.59	43.43

achieves the highest precision across all aggregated metrics, with 71.50 in the micro category, 64.20 in the macro category, 74.84 in weighted precision, and 50.16 in sample precision, underscoring the effectiveness of the full model setup.

Disabling the pose estimation strategy while keeping CDDR active results in a slight reduction in performance with regard to the whole model. Both precision and recall metrics are lower than using the whole framework in all types of combination. This results in a decrease of the F1-score, and underscores the significance of pose estimation in enhancing the model's accuracy.

Conversely, when CDDR is disabled and only pose estimation is used, results show mixed impacts. Although recall is slightly improved, precision decreases significantly with respect to the whole model. Combining both, the overall performance remains below that of the complete model, indicating that CDDR provides a valuable complementary effect.

Finally, when both CDDR and pose estimation are disabled, the model experiences a similar decline in performance with respect to the complete model. Although recall is improved with respect to other configurations, the precision and the F1-score are lower than those of the complete model setup in all averages, except for the sample-F1-score, which is similar.

These results suggest that incorporating the CDDR and pose modules slightly reduces recall but significantly improves precision. In other

Table 11
Ablation study performed on two stages: collision detection and duplicate removal (CDDR), and pose estimation for PPE matching (Pose).

CDDR	Pose	Micro			Macro			Weighted			Sample		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
✓	✓	**71.50**	68.15	**69.79**	**64.20**	63.99	**63.29**	**74.84**	68.15	**70.61**	**50.16**	55.09	50.57
✓	✗	70.10	66.81	68.42	63.13	63.02	62.28	73.26	66.81	69.17	48.05	53.78	48.67
✗	✓	66.90	70.47	68.64	60.27	65.91	62.45	68.85	70.47	69.19	48.68	57.28	50.44
✗	✗	66.96	**70.77**	68.81	60.17	**66.01**	62.45	68.92	**70.77**	69.37	48.87	**57.54**	**50.68**

words, these components filter out some correct detections which do not meet specific conditions, but the remaining instances provide a higher confidence. In the context of health and safety compliance and monitoring, it is crucial to ensure that workers labeled as safe are indeed wearing the appropriate attire.

In summary, the ablation study clearly demonstrates that the combination of CDDR and pose estimation provides the best overall performance. Each component uniquely enhances different aspects of the model's accuracy, and their synergistic effect is crucial for achieving the highest efficacy in PPE matching.

6. Conclusion

This paper presented an AI-based framework for automated PPE compliance monitoring in construction and industrial environments, aiming to enhance worker safety through deep learning techniques. The system integrates worker detection, pose estimation, and PPE recognition to address key challenges such as occlusions, shadows, and varying camera distances. Through extensive evaluations on different datasets — including the Zamakona and Pictor datasets, as well as additional construction site scenarios with diverse environmental and worker occupancy conditions — our framework demonstrates adaptability and robustness across different real-world settings.

The proposed approach incorporates a collision detection and duplicate removal module to handle complex cases involving overlapping or partially occluded workers, ensuring reliable PPE assignment. Additionally, pose estimation enhances worker-PPE matching, improving assignment accuracy in crowded or visually ambiguous environments. The ablation study confirms that these components contribute to the overall performance of the system, indicating their importance for effective compliance monitoring.

A comprehensive quantitative and qualitative evaluation against state-of-the-art methods, including the Xiong and Nath (A2 and A3) models, demonstrates the effectiveness of our approach. Our framework significantly outperforms Nath A2 and A3 (originally based on YOLOv3) in both datasets. To ensure a fair comparison, we employ an improved version of Nath A3, replacing YOLOv3 with InternImage for enhanced worker detection. While achieving comparable performance to the best Nath models in the Pictor dataset, our method consistently surpasses all alternatives in the Zamakona dataset. Additionally, compared to Xiong, our framework achieves higher recall and F1 scores, particularly improving PPE assignment in challenging conditions, demonstrating its robustness across different environments.

Qualitative assessments further confirm its generalization beyond training data, accurately identifying PPE compliance under varying lighting, occlusions, and worker densities. While detection performance is highest for helmets and vests, it decreases for smaller or less visible elements, such as gloves, particularly in long-distance scenarios. Nonetheless, the framework maintains a strong balance between precision and recall, making it a reliable solution for real-world applications.

Overall, this study presents a flexible framework for PPE compliance monitoring, addressing gaps in existing techniques. The proposed approach aims to contribute to future AI-driven safety enhancements, such as real-time alert mechanisms, compliance reporting, and risk analytics, thereby supporting automated safety monitoring in high-risk industrial settings.

CRediT authorship contribution statement

Leopoldo López: Writing – original draft, Validation, Software, Investigation, Conceptualization. **Jonay Suárez-Ramírez:** Writing – original draft, Validation, Software, Investigation, Formal analysis. **Miguel Alemán-Flores:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis. **Nelson Monzón:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nelson Monzon reports financial support was provided by Consejeria de Vicepresidencia Primera y de Obras Públicas, Infraestructuras, Transporte y Movilidad from Cabildo de Gran Canaria. Leopoldo López y Jonay Suárez-Ramírez reports financial support was provided by Qualitas Artificial Intelligence & Science. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is the result of a collaboration between the R&D company Qualitas Artificial Intelligence & Science (QAISC, www.qaisc.com) and the Imaging Technology Center (CTIM) at the University of Las Palmas de Gran Canaria, within the framework of research contract C2023/34, signed between the company and the Canarian Science and Technology Park Foundation of the University of Las Palmas de Gran Canaria.

It has also been supported by the Vicepresidencia Primera, Consejería de Obras Públicas, Infraestructuras, Transporte y Movilidad of the Cabildo de Gran Canaria, through the project referenced in Resolution "DETECCIÓN PRECISA IA".

We express our sincere thanks to Repnaval for providing access to the experimental shipyard environment necessary for this work, within the framework of the project "Proyecto de desarrollo y ensayo de un sistema de varada inteligente (SIVIS)", funded by the Spanish Ministry of Industry, Commerce and Tourism (I+D fondo de reestructuración naval, 2017–2020, ref. 5954/I+D-01). The visual data collected at the Repnaval facility were captured under internal protocol approval by the shipyard's management board and in full compliance with Spanish data protection laws (Ley Orgánica 3/2018, de Protección de Datos Personales y garantía de los derechos digitales). We are also grateful to the Oceanic Platform of the Canary Islands (PLOCAN) for providing visual data captured in the vicinity of the Port of Las Palmas de Gran Canaria, in accordance with the Memorandum of Understanding signed between QAISC and PLOCAN (PLOCAN CONSORCIO – REGISTRO DE SALIDA N° 670/21, dated December 22, 2021), which authorizes access to sensor data for research and innovation purposes.

Special thanks go to architect Néstor Ojeda Izquierdo (COACGC registration number 3317), for providing images derived from his professional architectural work in Gran Canaria. These photographs

were obtained in accordance with national and European data protection regulations, and without the inclusion of personally identifiable information.

All visual material included in this publication has been appropriately anonymized for dissemination (e.g., through face blurring techniques) to ensure that no individuals are identifiable. The usage of such data has been limited strictly to scientific and academic purposes, and all sources of image data are supported by signed legal certifications submitted as supplementary material.

We also acknowledge the valuable contribution of student Pablo Guilló Jiménez for his assistance in dataset cleaning and refinement.

Data availability

The authors do not have permission to share data.

References

[1] Y. Kang, S. Siddiqui, S.J. Suk, S. Chi, C. Kim, Trends of fall accidents in the US construction industry, J. Constr. Eng. Manag. 143 (8) (2017) 04017043, http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0001332, Publisher Copyright: © 2017 American Society of Civil Engineers..

[2] Occupational Safety and Health Administration, Worker safety series construction, 2022, pp. 1–36, URL https://www.osha.gov/Publications/OSHA3252/3252.html. (Accessed 6 July 2019).

[3] B.E. Mneymneh, M. Abbas, H. Khoury, Vision-based framework for intelligent monitoring of hardhat wearing on construction sites, J. Comput. Civ. Eng. (ISSN: 0887-3801) 33 (2) (2019) 04018066, http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000813, Article ID: 04018066, eissn: 1943-5487.

[4] J.-H. Lo, L.-K. Lin, C.-C. Hung, Real-time personal protective equipment compliance detection based on deep learning algorithm, Sustain. 15 (391) (2023) 391, http://dx.doi.org/10.3390/su15010391.

[5] A.M. Vukicevic, M. Petrovic, P. Milosevic, A. Peulic, K. Jovanovic, A. Novakovic, A systematic review of computer vision-based personal protective equipment compliance in industry practice: advancements, challenges and future directions, Artif. Intell. Rev. 57 (319) (2024) 1–28, http://dx.doi.org/10.1007/s10462-024-10978-x.

[6] A.M. Vukicevic, M. Djapan, V. Isailovic, D. Milasinovic, M. Savkovic, P. Milosevic, Generic compliance of industrial PPE by using deep learning techniques, Saf. Sci. 148 (2022) 105646, http://dx.doi.org/10.1016/j.ssci.2021.105646, URL https://www.sciencedirect.com/science/article/pii/S0925753521004860.

[7] R. Xiong, P. Tang, Pose guided anchoring for detecting proper use of personal protective equipment, Autom. Constr. 130 (2021) 103828, http://dx.doi.org/10.1016/j.autcon.2021.103828, URL https://www.sciencedirect.com/science/article/pii/S092658052100279X.

[8] 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, http://dx.doi.org/10.1016/j.autcon.2020.103085, URL https://www.sciencedirect.com/science/article/pii/S0926580519308325.

[9] J. Cheng, P.K.-Y. Wong, H. Luo, M. Wang, P.H. Leung, Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification, Autom. Constr. 139 (2022) 104312, http://dx.doi.org/10.1016/j.autcon.2022.104312, URL https://www.sciencedirect.com/science/article/pii/S0926580522001856.

[10] S. Kim, S.H. Hong, H. Kim, M. Lee, S. Hwang, Small object detection (SOD) system for comprehensive construction site safety monitoring, Autom. Constr. 156 (2023) 105103, http://dx.doi.org/10.1016/j.autcon.2023.105103, URL https://www.sciencedirect.com/science/article/pii/S0926580523003631.

[11] B.E. Mneymneh, M. Abbas, H. Khoury, Automated hardhat detection for construction safety applications, Procedia Eng. 196 (2017) 895–902, http://dx.doi.org/10.1016/j.proeng.2017.08.022, URL https://www.sciencedirect.com/science/article/pii/S1877705817331430, Creative Construction Conference.

[12] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T.M. Rose, W. An, Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Autom. Constr. 85 (2018) 1–9, http://dx.doi.org/10.1016/j.autcon.2017.09.018, URL https://www.sciencedirect.com/science/article/pii/S0926580517304429.

[13] Z. Xie, H. Liu, Z. Li, Y. He, A convolutional neural network based approach towards real-time hard hat detection, IEEE Int. Conf. Prog. Inform. Comput. (2018) 430–434, URL https://api.semanticscholar.org/CorpusID:146119069.

[14] 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, http://dx.doi.org/10.1016/j.autcon.2019.102894, URL https://www.sciencedirect.com/science/article/pii/S092658051930264X.

[15] X. Wu, Y. Li, J. Long, S. Zhang, S. Wan, S. Mei, A remote-vision-based safety helmet and harness monitoring system based on attribute knowledge modeling, Remote. Sens. 15 (2) (2023) 347, http://dx.doi.org/10.3390/rs15020347, URL https://www.mdpi.com/2072-4292/15/2/347.

[16] M. Yang, Z. Yang, Y. Guo, S. Su, Z. Fan, A novel YOLO based safety helmet detection in intelligent construction platform, in: Intelligent Equipment, Robots, and Vehicles, Springer Singapore, Singapore, ISBN: 978-981-16-7213-2, 2021, pp. 268–275, http://dx.doi.org/10.1007/978-981-16-7213-2_26.

[17] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, Y. Zhao, Fast personal protective equipment detection for real construction sites using deep learning approaches, Sensors 21 (10) (2021) 3478, http://dx.doi.org/10.3390/s21103478, URL https://www.mdpi.com/1424-8220/21/10/3478.

[18] M. Zhao, M. Barati, Substation safety awareness intelligent model: Fast personal protective equipment detection using GNN approach, IEEE Trans. Ind. Appl. 59 (3) (2023) 3142–3150, http://dx.doi.org/10.1109/TIA.2023.3234515.

[19] H. Seong, H. Choi, H. Cho, S. Lee, H. Son, C. Kim, Vision-based safety vest detection in a construction scene, in: M.-Y. Cheng, H.-M. Chen, K.C. Chiu (Eds.), Proceedings of the 34th International Symposium on Automation and Robotics in Construction, Tribun EU, s.r.o., Brno, Taipei, Taiwan, 2017, pp. 288–293, http://dx.doi.org/10.22260/ISARC2017/0039.

[20] X. Wang, D. Niu, P. Luo, C. Zhu, L. Ding, K. Huang, A safety helmet and protective clothing detection method based on improved-yolo v 3, in: 2020 Chinese Automation Congress, 2020, pp. 5437–5441, http://dx.doi.org/10.1109/CAC51589.2020.9327187.

[21] Z. Wang, Z. Cai, Y. Wu, An improved YOLOX approach for low-light and small object detection: PPE on tunnel construction sites, J. Comput. Des. Eng. 10 (3) (2023) 1158–1175, http://dx.doi.org/10.1093/jcde/qwad042, arXiv:https://academic.oup.com/jcde/article-pdf/10/3/1158/50781123/qwad042.pdf.

[22] R.N. Bhadeshiya, K.N. Brahmbhatt, J.R. Pitroda, Hard-hat detection using YOLOv4, in: 2021 Second International Conference on Electronics and Sustainable Communication Systems, 2021, pp. 1114–1120, http://dx.doi.org/10.1109/ICESC51422.2021.9532896.

[23] V.S.K. Delhi, R. Sankarlal, A. Thomas, Detection of personal protective equipment (PPE) compliance on construction site using computer vision based deep learning techniques, Front. Built Environ. 6 (2020) 1–10, http://dx.doi.org/10.3389/fbuil.2020.00136, URL https://www.frontiersin.org/articles/10.3389/fbuil.2020.00136.

[24] J. Lee, S. Lee, Construction site safety management: A computer vision and deep learning approach, Sensors 23 (2) (2023) 944, http://dx.doi.org/10.3390/s23020944, URL https://www.mdpi.com/1424-8220/23/2/944.

[25] J. Kim, S. Chi, J. Kim, 3D pose estimation and localization of construction equipment from single camera images by virtual model integration, Adv. Eng. Inform. 57 (2023) 102092, http://dx.doi.org/10.1016/j.aei.2023.102092, URL https://www.sciencedirect.com/science/article/pii/S1474034623002203.

[26] X. Zhou, S. Li, J. Liu, Z. Wu, Y.F. Chen, Construction activity analysis of workers based on human posture estimation information, Eng. 33 (2024) 225–236, http://dx.doi.org/10.1016/j.eng.2023.10.004, URL https://www.sciencedirect.com/science/article/pii/S2095809923004691.

[27] A. Moohialdin, F. Lamari, M. Marc, B. Trigunarsyah, A real-time computer vision system for workers' PPE and posture detection in actual construction site environment, in: C.M. Wang, V. Dao, S. Kitipornchai (Eds.), EASEC16, Springer Singapore, Singapore, ISBN: 978-981-15-8079-6, 2021, pp. 2169–2181, http://dx.doi.org/10.1007/978-981-15-8079-6_199.

[28] A. Kelm, L. Laußat, A. Meins-Becker, D. Platz, M.J. Khazaee, A.M. Costin, M. Helmus, J. Teizer, 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, http://dx.doi.org/10.1016/j.autcon.2013.08.009, URL https://www.sciencedirect.com/science/article/pii/S0926580513001234.

[29] D. Wang, W. Li, X. Liu, N. Li, C. Zhang, UAV environmental perception and autonomous obstacle avoidance: A deep learning and depth camera combined solution, Comput. Electron. Agric. 175 (2020) 105523, http://dx.doi.org/10.1016/j.compag.2020.105523, URL https://www.sciencedirect.com/science/article/pii/S0168169920303379.

[30] H. Luo, M. Wang, P.K.-Y. Wong, J.C. Cheng, Full body pose estimation of construction equipment using computer vision and deep learning techniques, Autom. Constr. 110 (2020) 103016, http://dx.doi.org/10.1016/j.autcon.2019.103016, URL https://www.sciencedirect.com/science/article/pii/S092658051930634X.

[31] H. Son, H. Seong, H. Choi, C. Kim, Real-time vision-based warning system for prevention of collisions between workers and heavy equipment, J. Comput. Civ. Eng. 33 (5) (2019) 04019029, http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000845, arXiv:https://ascelibrary.org/doi/pdf/10.1061/%28ASCE%29CP.1943-5487.0000845, URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CP.1943-5487.0000845.

[32] A. Akinsemoyin, I. Awolusi, D. Chakraborty, A.J. Al-Bayati, A. Akanmu, Unmanned aerial systems and deep learning for safety and health activity monitoring on construction sites, Sensors 23 (15) (2023) 6690, http://dx.doi.org/10.3390/s23156690, URL https://www.mdpi.com/1424-8220/23/15/6690.

[33] S.S. Megha Nain, S. Chaurasia, Authentication control system for the efficient detection of hard-hats using deep learning algorithms, J. Discret. Math. Sci. Cryptogr. 24 (8) (2021) 2291–2306, http://dx.doi.org/10.1080/09720529.2021.2011109, arXiv:https://doi.org/10.1080/09720529.2021.2011109.

[34] Y. Zhang, J. Shi, D. Wang, C. Pang, Z. Yang, B. Wu, Y. Xu, W. Du, Detection on safety helmet wearing of distribution network construction based on YOLOv5-btri algorithm, in: 2022 2nd Asia-Pacific Conference on Communications Technology and Computer Science, 2022, pp. 517–524, http://dx.doi.org/10.1109/ACCTCS53867.2022.00110.

[35] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 7464–7475, http://dx.doi.org/10.1109/CVPR52729.2023.00721.

[36] Y. Xu, J. Zhang, Q. Zhang, D. Tao, ViTPose: Simple vision transformer baselines for human pose estimation, in: Advances in Neural Information Processing Systems, 2022, pp. 38571–38584, URL https://dl.acm.org/doi/10.5555/3600270.3603065.

[37] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: deformable transformers for end-to-end object detection, 2020, pp. 1–14, CoRR abs/2010.04159, arXiv:2010.04159.URL https://arxiv.org/abs/2010.04159.

[38] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L.M. Ni, H.-Y. Shum, DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection, 2022, arXiv:2203.03605, URL https://arxiv.org/abs/2203.03605.

[39] S. Zhang, X. Wang, J. Wang, J. Pang, C. Lyu, W. Zhang, P. Luo, K. Chen, Dense distinct query for end-to-end object detection, 2023, arXiv:2303.12776, URL https://arxiv.org/abs/2303.12776.

[40] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3349–3364, http://dx.doi.org/10.1109/TPAMI.2020.2983686.

[41] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, S.-T. Xia, Simcc: A simple coordinate classification perspective for human pose estimation, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham, ISBN: 978-3-031-20068-7, 2022, pp. 89–106, http://dx.doi.org/10.48550/arXiv.2107.03332.

[42] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, Y. Qiao, InternImage: Exploring large-scale vision foundation models with deformable convolutions, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 14408–14419, http://dx.doi.org/10.1109/CVPR52729.2023.01385.