sustainability

MDPI

Article

Real-Time Personal Protective Equipment Compliance Detection Based on Deep Learning Algorithm

Jye-Hwang Lo *, Lee-Kuo Lin and Chu-Chun Hung

Department of Civil Engineering, National Taipei University of Technology, Taipei City 106, Taiwan
* Correspondence: tommy70820@gmail.com

Abstract: The construction industry is one of the most dangerous industries in the world due to workers being vulnerable to accidents, injuries and even death. Therefore, how to effectively manage the appropriate usage of personal protective equipment (PPE) is an important research issue. In this study, deep learning is applied to the PPE inspection model to verify whether construction workers are equipped in accordance with the regulations, and this is expected to reduce the probability of related occupational disasters caused by the inappropriate use of PPE. The method is based on the YOLOv3, YOLOv4 and YOLOv7 algorithms to detect worker's helmets and high-visibility vests from images or videos in real time. The model was trained on a new PPE dataset collected and organized by this study; the dataset contains 11,000 images and 88,725 labels. According to the test results, can achieve a 97% mean average precision (mAP) and 25 frames per second (FPS). The research result shows that the detection and counting data in this method have performed well and can be applied to the real-time PPE detection of workers at the construction job site.

Keywords: personal protective equipment; deep learning; object detection; YOLOv3; YOLOv4; YOLOv7

Citation: Lo, J.-H.; Lin, L.-K.; Hung, C.-C. Real-Time Personal Protective Equipment Compliance Detection Based on Deep Learning Algorithm. *Sustainability* **2023**, *15*, 391. https://doi.org/10.3390/su15010391

Academic Editors: Pin-Chao Liao and Tingya Hsieh

Received: 8 November 2022
Revised: 2 December 2022
Accepted: 22 December 2022
Published: 26 December 2022

1. Introduction

The construction industry has one of the most dangerous occupational injuries in the civil construction business. The situation is caused by risky duty content and a dangerous working environment. Common accidents such as falling, being hit by collapsed objects and collisions could lead to injuries and deaths. Throughout the years, authorities have formulated and imposed various safety regulations and on-site inspections to ensure a safer working environment, but still, the number of casualties remains as high despite the new regulations.

According to the statistics of the Ministry of Labor, Taiwan, the incidence rate of fatal occupational injuries in the construction industry remains higher than others in 2020, with up to 71% of major occupational injuries caused by unsafe equipment [1]. The U.S. Occupational Safety and Health Administration (OSHA) has also indicated that the most frequent occupational injuries and deaths can be effectively prevented if workers wear appropriate personal protective equipment (PPE): helmets, safety vests, steel toe boots, goggles and gloves [2]; particularly, helmet-wearing can significantly reduce the impacts caused by falling objects, electrocution due to hanging cables and other common occupational injuries. Moreover, demanding construction workers to wear high-visibility safety vests can effectively prevent heavy construction equipment from hitting them; therefore, how to efficiently and properly manage the use of personal protective equipment (PPE) still deserves more in-depth study and discussion.

The environment of construction operations changes rapidly, and appropriate monitoring systems can help accurately understand the information of the construction site in real time, thus allowing the inspector to rapidly respond to safety issues. The traditional method of monitoring PPE still mainly consists of manual inspection and sensors. Visual

inspection is very time-consuming; thus, its effectiveness depends largely on the safety knowledge and experience of the inspector, which makes it prone to omission or prejudices. Although the existing sensor detection methods can immediately transmit important monitoring information, their installation and maintenance costs are still very high, which may affect their practicality on the construction site; therefore, it is still necessary to find efficient ways to automate the monitoring process to reduce the risk of injury and death and to ensure the safety of the construction environment.

In recent years, with the development of artificial intelligence (AI) and deep learning, various applications of image recognition have become more and more common. If the technology can be applied to analyze the images recorded by construction site surveillance cameras to verify whether the construction workers are wearing PPE or not, it will provide a faster, more accurate and comprehensive understanding of the construction site conditions without excessive additional cost [3].

To summarize, this study contains the following contributions:

1. This study built a new large-scale PPE image dataset, including 11,000 images for detecting labels of hard hats and high-visibility vests. The PPE dataset was released for the first time in this study and we provided the access method at the end of the article;
2. This study developed a YOLOv7 [4]-based PPE compliance detection model and compared its performance with YOLOv3 [5] and YOLOv4 [6] models in complex environments.

2. Literature Review

To implement the analysis procedure of this study, this research analyzes and compiles related relevant literature, including personal protective equipment (PPE) detection and deep learning-related technologies, which will be discussed in the following section.

Object detection is the process of identifying objects in an image and accurately positioning their position in the image. Deep learning technology has become a powerful solution for learning features directly from data and has made a breakthrough in the field of general object detection [7]. Most of the current mainstream object detection algorithms are still improved based on the convolutional neural network (CNN) architecture, e.g., Regions with CNN features(R-CNN) [8] combines region proposals with CNN, and when the marked training data are too small, pre-training can be carried out under the supervision of auxiliary tasks, and the performance can be improved by fine-tuning specific areas. Although R-CNN [8] has a certain detection quality, it still has the issue of slowness, so the fast region-based convolutional network (Fast R-CNN) method [9] is derived, by developing a simplified training process to enable the end-to-end detector for training, and learning the softmax function classifier and specific category bounding box regression at the same time of training. A single training process can update all network layers, and the feature cache does not require storage space, so the detection quality is higher. Compared with R-CNN [8], the training speed is about three times faster, and the test speed is about 10 times faster. Although Fast R-CNN [9] speeds up the detection process, Ren et al. are still attempting to break the bottleneck of region proposal computation, and the proposed faster region-based convolutional neural network (Faster R-CNN) method [10] uses the region proposal network (RPN) to simultaneously detect the object range and score of each position. In order to solve the problem of pixel-level object instance segmentation, He et al. proposed the mask region-based convolutional neural network (Mask R-CNN) [11] by extending Faster R-CNN [10]. However, although the above algorithm has a certain accuracy, it still cannot achieve the effect of real-time object detection. At present, the most popular algorithms in the real-time object detection field are object detection via region-based fully convolutional networks (R-FCNs) [12], single-shot multibox detector (SSD) [13], you only look once (YOLO) [14] series algorithms, and other methods.

In the construction industry, traditional safety management approaches, such as training and preventative controls, are inadequate to protect construction workers because

it is hard to fully predict the occurrence of occupational injuries during the design and planning phase, and automatic monitoring technology can not only facilitate real-time on-site monitoring but also save labor cost and improve site safety. So far, the research on PPE detection can be categorized into sensor-based and vision-based detection methods.

The sensor-based techniques generally perform PPE detection by installing sensors and analyzing their signals, e.g., Zigbee [15], radio frequency identification (RFID) [16], cyber–physical systems (CPSs) developed by using Zigbee and RFID [17], the real-time location system (RTLS) with integrated pressure sensor [18] and the Internet of Things (IoT)-based architecture including infrared beam detectors and thermal infrared sensors [19].

Generally, the current sensor-based detection and monitoring technology is further restricted by the fact that each construction worker must wear a tag or sensor, which requires a complicated training process, and it is difficult to correctly identify whether everyone on the construction site is wearing a PPE, and the actual use of tags or sensors will incur a lot of costs. In addition, sensors may also be accompanied by personal positioning privacy issues and may generate health concerns such as electromagnetic fields, which makes some construction workers unwilling to wear sensors for a long time.

Compared with sensor-based detection methods, due to the advancement of computer vision and image recognition technology, vision-based PPE detection technology has received more and more attention. For example, Hao and Zhao proposed a safety helmet identification method based on K-nearest neighbors (KNNs), histograms of oriented gradients (HOGs) and hierarchical support vector machines (H-SVMs) [20]. Xie et al. proposed a CNN-based approach toward hard hat detection (CAHD) to detect hard hats [21]. Fang et al. used Faster R-CNN [10] to detect non-hard-hat-use (NHU) [3] and Wu et al. used single-shot multibox detector (SSD) to detect NHU [22]. Nath et al. presents three YOLOv3 [5]-based models, and the best performing model achieves 72.3% mAP, and can process 11 FPS on a laptop computer while simultaneously detecting individual workers and verifying PPE compliance [23].

Accordingly, most of the PPE detection methods based on computer vision are focused on detecting a single PPE object (e.g., helmet), and some of the methods have not been able to achieve the effect of immediate detection. Although the detection methods based on computer vision have been extensively researched, there are still situations where it is difficult to accurately detect individual images at a certain distance from the camera, and it is difficult to extend the study to other construction sites because the construction workers and environmental background characteristics are different. In addition, if there are too many construction workers in the same image area, it will lead to the partial occlusion of each other, which also makes the detection of PPE more difficult. There are still a few studies that focus on the function of deep learning algorithms to identify multiple types of PPE, and as the number and types of PPE increase, the complexity of detection tasks will grow significantly, which is still challenging.

3. Materials and Methods

The processor used in the study was an Intel® Core™ i5-10400F, with a 2.8 GHz processor, 32 GB RAM, 500 GB Solid-state drive and 11 GB NVIDIA GTX 1080Ti GPU. All the implementation was performed by using the Python programming language under Windows 10, and the deep learning framework is TensorFlow and Keras.

3.1. YOLO Networks

The you only look once (YOLO) algorithm [14] converts the problem of target detection into a regression problem, which predicts the classes and bounding boxes of the whole image at once. Consequently, the detection speed and mAP are better than the Faster R-CNN [10,24], but there is still the problem of the low accuracy in detecting small objects and the limitation of the classification to a single class of objects [25]. To improve the YOLO algorithm [14], YOLOv2 [26] and YOLO9000 [26] have imported the multi-scale training method and DarkNet19 architecture. YOLO9000 [26] is a real-time framework for

detecting over 9,000 categories of objects through the joint optimization of detection and classification. YOLOv3 [5] is an object detector evolved from YOLO [14] and YOLOv2 [26]. YOLOv3 [5] uses Darknet-53 for performing feature extractions and is much more powerful than Darknet19. Compared to earlier versions, YOLOv3 [5] not only has high accuracy and detection speed, but also works well with small target detection. The structure of YOLOv3 [5] is shown in Figure 1. The input image is down-sampled 5 times by YOLOv3 [5], and the predictions are made in the last 3 down samplings. The last 3 down samplings include three-scale target detection feature maps. The large and small feature maps provide the location and deep semantic information, respectively, and the YOLOv3 [5] network algorithm is shown in Figure 2.

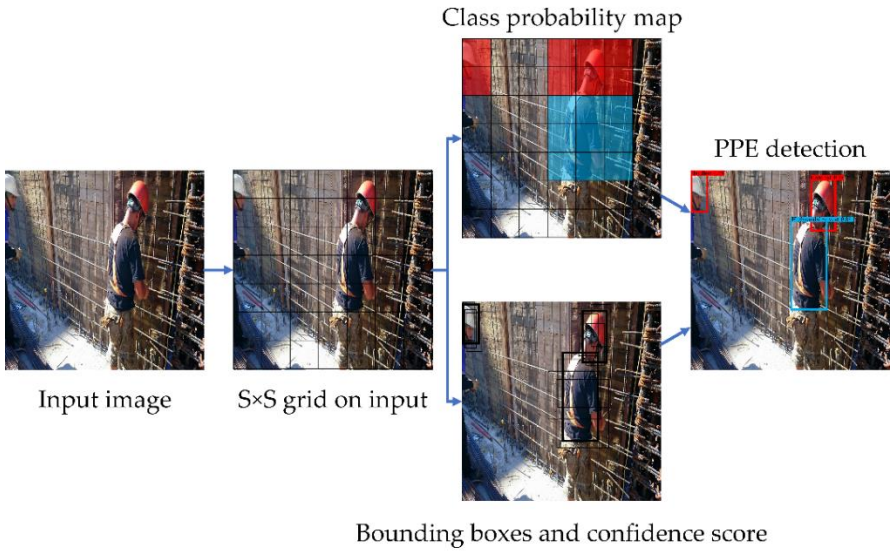


Figure 1. YOLOv3 network detection architecture.

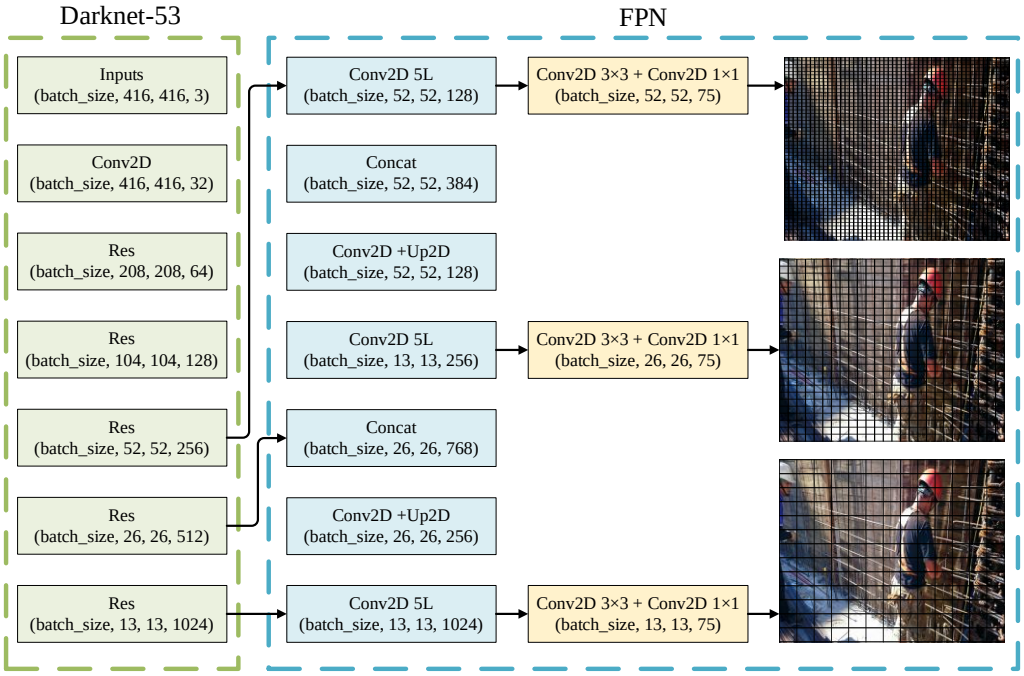


Figure 2. YOLOv3 network algorithm.

The loss function (Equations (1) and (2)) of YOLOv3 [5] consists of localization loss, confidence loss and classification loss.

$$\begin{aligned}\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{i,j}^{obj} \left[(t_x - \hat{t_x})^2 + (t_y - \hat{t_y})^2 + (t_w - \hat{t_w})^2 + (t_h - \hat{t_h})^2 \right] \\ + \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{i,j}^{obj} \left[-log(p_c) + \sum_{i=1}^{n} BCE(\hat{c_i}, c_i) \right] \\ + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{i,j}^{noobj} [-log(1 - p_c)]\end{aligned} \tag{1}$$

where $\lambda_{coord} = 5$, $\lambda_{noobj} = 0.5$, S is the number of grids and B is the number of anchors. $\mathbb{I}_{i,j}^{obj}$ indicates whether there is an object in the i-th cell, and the definition of $\mathbb{I}_{ij}^{noobj}$ is the opposite of $\mathbb{I}_{i,j}^{obj}$. $(t_x,\ t_y)$ is the center of the prediction bounding box, $(\hat{t_x}, \hat{t_y})$ is the center of the ground truth. $(t_w,\ t_h)$ is the scale factor for resizing the bounding box, $(\hat{t_w},\ \hat{t_h})$ is the correct scale factor. p_c is the confidence of the grid. n is the number of classes.

$$BCE(\hat{c_i}, c_i) = -\hat{c_i} log(c_i) - (1 - \hat{c_i}) log(1 - c_i) \tag{2}$$

where $\hat{c_i}$ and c_i represent the actual and predicted classes, respectively.

Currently, YOLOv4 [6] is one of the most popular object detection methods based on deep learning with high speed and precision. YOLOv4 [6] provides an advanced detector that is faster and more accurate than all other available solutions [27]. In order to further develop a more powerful object detection model, YOLOv4 [6] creates a new feature extractor backbone called CSPDarknet53, and imports two methods based on existing algorithms. One is the "Bag-of-Freebies" [28], which can improve the accuracy of the object detector without raising the cost of inference, and the other is the "Bag-of-Specials", which significantly improves the accuracy of object detection by slightly increasing the cost of inference. YOLOv4 [6] also greatly improves the speed and accuracy of small object detection through major innovations such as Mosaic data augmentation, self-adversarial training (SAT), cross mini-batch normalization (CmBN), modified spatial attention module (SAM) and modified path aggregation network (PAN). To confirm the impact of the enhanced algorithm on small object detection, a more challenging 416 × 416 resolution was selected for the experiment, and the YOLOv4 [6] structure is shown in Figure 3. With the inputted image size of 416 × 416, the proposed model can predict bounding boxes at the detector head at three different scales with better performance and increased accuracy: 52 × 52 × 255, 26 × 26 × 255 and 13 × 13 × 255, which are applicable for detecting small-scale, medium-scale and large-scale objects, respectively. Although the original anchor of YOLOv4 [6] corresponds to various objects and sizes, it is always challenging to meet the needs of specific applications, particularly for applications that require appropriate image. The 9 parameters of the prior bounding box in the original network corresponding to the detection of small, medium and large objects may not meet the practical requirements of PPE detection. Therefore, the parameters of the prior bounding box are readjusted by using the k-means clustering algorithm to improve the correspondence between the bounding box and the object.

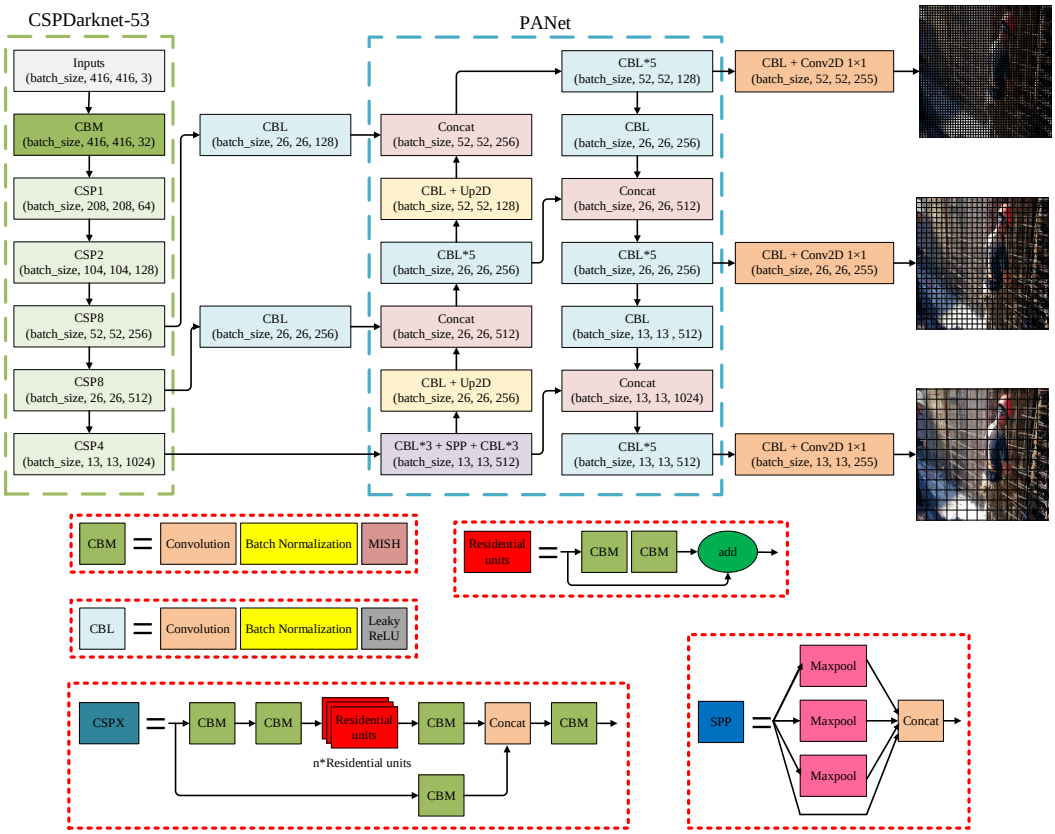


Figure 3. YOLOv4 network algorithm.

The loss function of the YOLOv4 [6] model uses the same classification loss and confidence loss as the YOLOv3 [5] model, but complete intersection over union (CIoU) (Equations (3)–(5)) is used as the replacement for mean squared error (MSE).

$$Loss_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{3}$$

where $\rho^2(b, b^{gt})$ is the distance between the center points of the prediction box and the ground truth, and c is the diagonal length of the minimum enclosing box covering the two boxes. IoU represents the intersection and union ratio between predicted bounding box and ground truth bounding box.

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{4}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \tag{5}$$

where w^{gt} and h^{gt} are the width and height of the ground truth bounding box, respectively, while w and h represent the width and height of the prediction bounding box, respectively.

YOLOv7 [4] is the most recent work of the YOLO series, and the YOLOv7 [4] structure is shown in Figure 4. Wang et al. proposed YOLOv7 [4] in 2022 and claimed that it is the state-of-the-art algorithm for real-time object detection with higher detection accuracy and faster inference speed. In comparison with YOLOv4 [6], YOLOv7 [4] has 75% less parameters, 36% less computation and brings 1.5% more average precision (AP) [29]. The speed and accuracy exceed all known object detectors in the range of 5–160 FPS. The trainable bag-of-freebies proposed by YOLOv7 [4] focuses more on optimizing the training process while improving object detection accuracy, but without inflating the cost of inference. The proposed extended ELAN (E-ELAN) uses expand, shuffle and merge

cardinality to repeatedly improve network learning capability without destroying the original gradient path and enhance the features learned from different feature maps. The paper also proposes a compound scaling method to improve the AP with less parameters and computation more efficiently. However, there are very few studies on the application of the YOLOv7 [4] model for PPE detection at present.

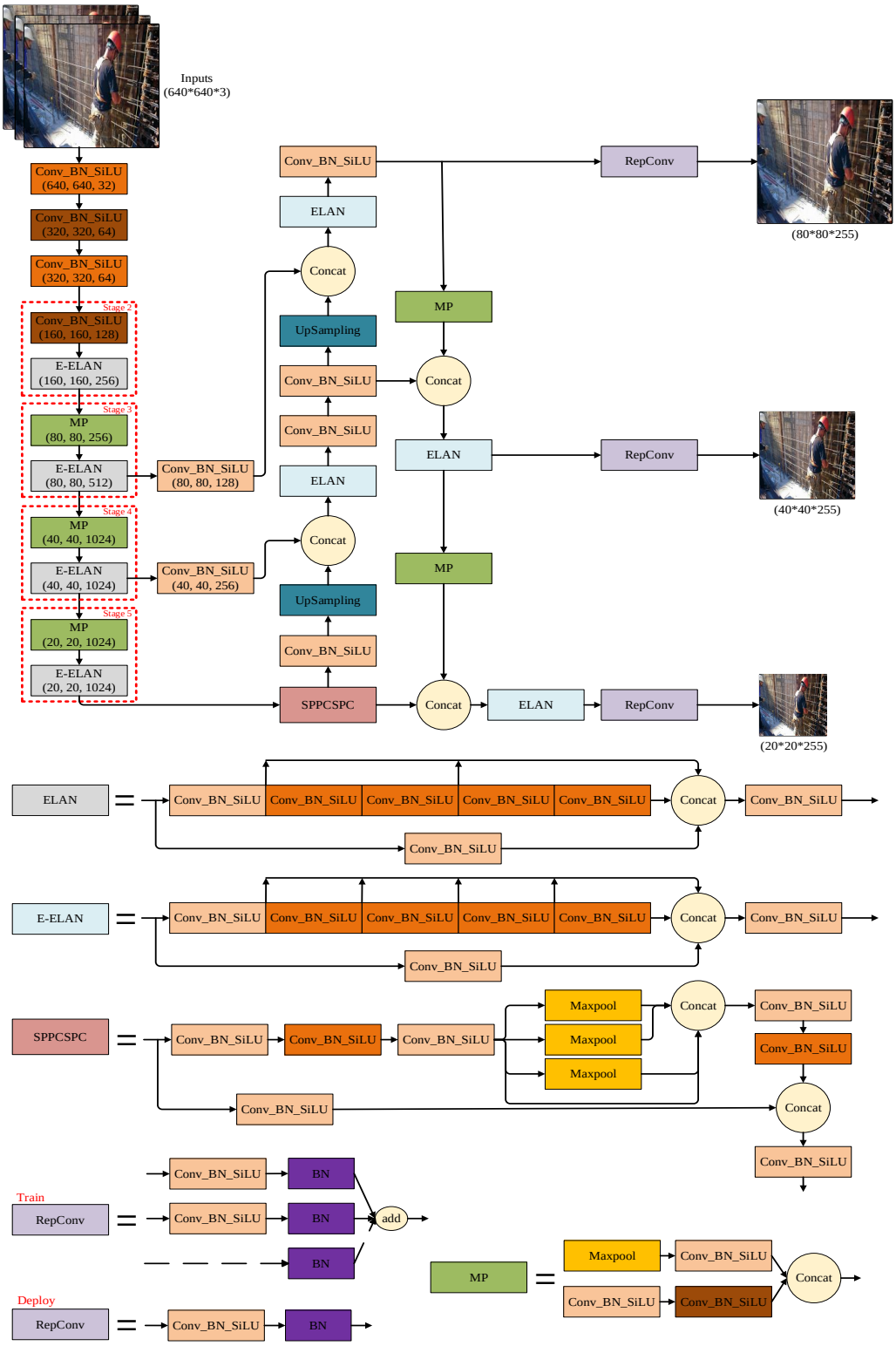


Figure 4. YOLOv7 network algorithm.

3.2. Image Annotation and Dataset Development

To ensure the diversity and complexity of the developed dataset, a large number of images of construction workers from different types of work and different construction environments will be captured by cameras and Web mining [30], and there are no other results for this dataset. The open-source LabelImg script (https://github.com/heartexlabs/labelImg) (accessed on 1 January 2022) from GitHub was used for annotating the dataset. It is a bounding box that explicitly annotates objects in images, as shown in Figure 5. It has been developed to provide a sample for machine learning. After launching the LabelImg [31] script, the target samples in each image were tagged. Subsequently, an XML file containing the target type and coordinates required for the dataset training was generated. The new PPE dataset contains 11,000 images and 88,725 PPE labels for various construction environments, as shown in Table 1 and Figure 6.

Figure 5. Sample image annotation using the LabelImg tool.

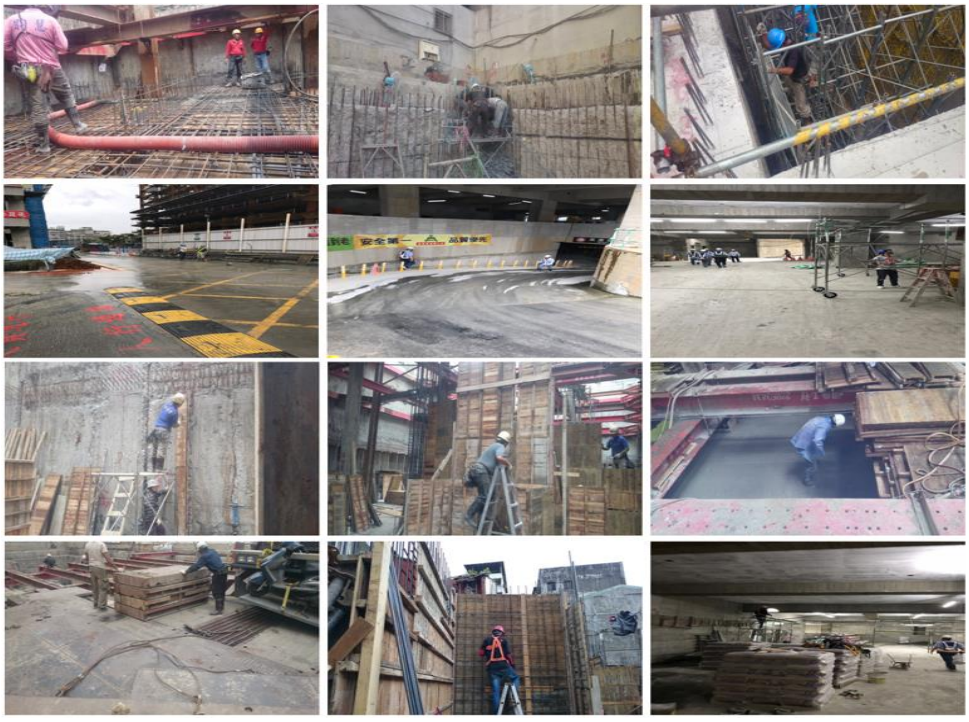


Figure 6. The PPE dataset.

Table 1. Labels and classes of PPE.

Hard Hat	No Hard Hat	High-Visibility Vest	NO High-Visibility Vest
27,905	26,163	12,197	22,460

3.3. Data Pre-Processing

Data pre-processing is a critical step in data mining [32]. The proper execution of the pre-processing steps effectively limits data irregularities and improves data quality for reliable further analysis. The data cleaning steps cover missing data, mixed formats and duplicate or incorrect records. Different data cleaning tasks focus on different types of errors [33]. In this study, many inadequate training data were eliminated and data augmentation was used to solve the problem of limited data and overfitting. These augmentations artificially inflate the size of the training dataset by either data warping or oversampling [34]. For classification tasks, the robustness of the database is subjected to improvement by simply expanding the data in various ways, such as rotation, flip and size adjustment. However, in order to ensure the positioning of the bounding box on the image during object detection, it is necessary to change the position of the bounding box simultaneously during the geometry computation. As shown in Figure 7, during the YOLOv4 [6] and YOLOv7 [4] training, each step has a 50% chance of using Mosaic data augmentation, and the Mosaic-enhanced images also have a 50% chance of using the mixup [35] to alleviate adversarial perturbation. This study attempts to create a better data enhancement strategy close to the actual situation. Since the images of construction workers wearing PPE may be affected by illumination, posture and object occlusion, various data enhancement techniques such as flipping, cropping, noise injection and color space transformations will be applied randomly throughout the training stage to improve the dataset and enhance the simulation under different circumstances.

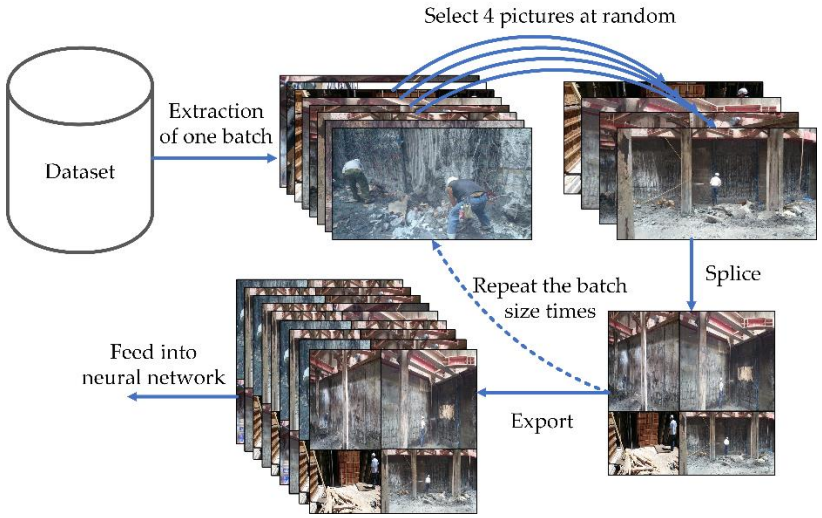


Figure 7. Mosaic data augmentation.

4. Results

4.1. Evaluation Metrics

In order to evaluate the performance of different models, the neural network predictions were judged by the mAP, per-class AP [29], precision (Equation (6)), recall (Equation (7)) and F1 score (Equation (8)). First, we computed the AP (Equation (9)) as the area under the precision/recall curve by numerical integration while the mAP (Equation (10)) is the AP averaged over four classes. Additionally, we make sure the IoU (Intersection over Union) (Equation (11)) exceeds 50%. The schematic diagram of the IoU calculation is shown in Figure 8.

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (6)$$

$$\text{recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (7)$$

$$\text{F1} = (2 \cdot \text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall}) \quad (8)$$

where TP are true positive, FP are false positive and FN are false negative classifications.

$$\mathrm{mAP} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=0}^{\mathrm{N}} \mathrm{APi} \tag{9}$$

where APi is the AP in the ith class and N is the total number of classes being evaluated.

$$\mathrm{AP} = \int_0^1 \mathrm{Precision} \cdot \mathrm{Recalldr} \tag{10}$$

where r represents the integral variable, which is used to determine the integral of precision*recall and is between 0 and 1.

$$\mathrm{IoU} = (\mathrm{Area}(\mathrm{Bp} \cap \mathrm{Bg}))/(\mathrm{Area}(\mathrm{Bp} \cup \mathrm{Bg})) \tag{11}$$

where Bp∩Bg and Bp∪Bg indicate, respectively, the intersection and union of the predicted and ground-truth bounding boxes.

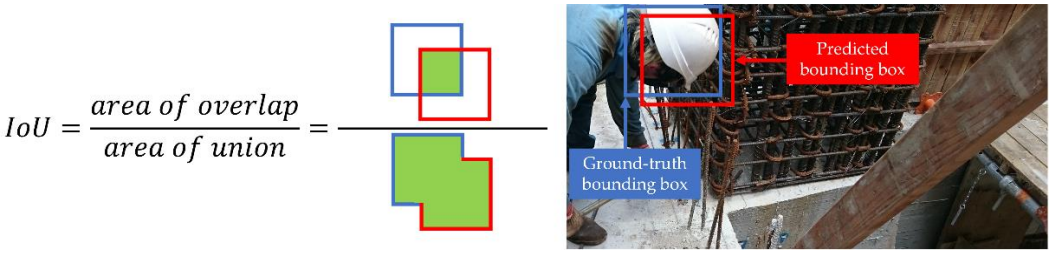


Figure 8. IoU.

4.2. Experimental Results and Analysis

Table 2 presents four parameters: average precision, precision, recall and F1 score to evaluate the predictive performance of the three models.

Table 2. Performances of the developed PPE detection models.

	Average Precision	Precision	Recall	F1 Score
YOLOv3- All	87.46%	94.18%	78.46%	85.39%
YOLOv3- Hard hat	92.75%	91.55%	88.68%	90.09%
YOLOv3- No hard hat	85.53%	97.96%	76.03%	85.61%
YOLOv3- High-visibility vest	86.50%	93.15%	78.90%	85.43%
YOLOv3- NO High-visibility vest	85.06%	94.07%	70.24%	80.43%
YOLOv4- All	91.16%	90.50%	92.74%	91.59%
YOLOv4- Hard hat	93.89%	90.43%	95.29%	92.80%
YOLOv4- No hard hat	94.42%	91.08%	95.54%	93.26%
YOLOv4- High-visibility vest	88.51%	91.00%	90.18%	90.59%
YOLOv4- NO High-visibility vest	87.82%	89.49%	89.95%	89.72%
YOLOv7- All	97.29%	91.75%	96.39%	94.01%
YOLOv7- Hard hat	97.95%	92.25%	98.59%	95.31%
YOLOv7- No hard hat	97.54%	92.75%	96.69%	94.68%
YOLOv7- High-visibility vest	97.16%	91.15%	95.79%	93.41%
YOLOv7- NO High-visibility vest	96.49%	90.83%	94.49%	92.62%

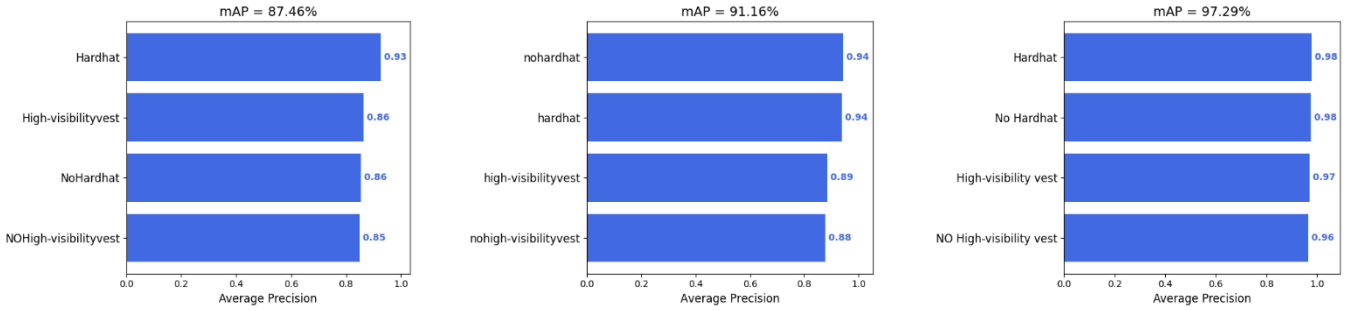
In the column of average precision, the YOLOv7 [4] model obtains the greatest average precision values among Hard hat, No hard hat, High-visibility vest and NO High-visibility vest, which are 97.95%, 97.54%, 97.16% and 96.49%, respectively. More precisely, the corresponding average precision values of the YOLOv7 [4] model are 5.20%, 12.01%, 10.66% and 11.43%, respectively, higher than YOLOv3 [5] model and 4.06%, 3.12%, 8.65% and 8.67%, respectively, higher than YOLOv4 [6] model.

In the column of precision, the YOLOv3 [5] model achieves the greatest precision value in all objects, which are 3.68% and 2.43% higher than the value of the YOLOv4 [6] model and the YOLOv7 [4] model, respectively. Additionally, the YOLOv3 [5] model has the highest precision in the No Hard hat, High-visibility vest and NO High-visibility vest objects, which are 6.88%, 2.15% and 4.58% higher than the YOLOv4 [6] model and 5.21%, 2.00% and 3.24% higher than the YOLOv7 [4] model, respectively. The YOLOv7 [4] model has the greatest precision in the Hardhat object, which is over 92%.

Regarding the column of recall, the YOLOv7 [4] model achieves the greatest recall values of 98.59%, 96.69%, 95.79% and 94.49% for the Hard hat, No Hard hat, High-visibility vest and NO High-visibility vest objects. More precisely, the corresponding recalls of the YOLOv7 [4] model are 9.91%, 20.66%, 16.89% and 24.25% higher than the YOLOv3 [5] model and 3.30%, 1.15%, 5.61% and 4.54% higher than the YOLOv4 [6] model.

Concentrating on the column of F1 score, the YOLOv7 [4] model still surpasses the other models. It achieves the greatest F1 score among the Hard hat, No Hard hat, High-visibility vest and NO High-visibility vest objects, which are 95.31%, 94.68%, 93.41% and 92.62%. Correspondingly, they are 5.22%, 9.07%, 7.98% and 12.19% higher compared to the YOLOv3 [5] model and 2.51%, 1.42%, 2.82% and 2.90% higher than the YOLOv4 [6] model.

From the result of the mAP presented in Figure 9, the YOLOv7 [4] model has outperformed the other models in accuracy and detection speed, and the model can achieve 97.29 mAP and 25.02 FPS. Overall, the YOLOv7 [4] model performs better than the YOLOv3 [5] model and the YOLOv4 [6] model, except for the fact that the performance of the precision value is slightly lower.



(**a**) (**b**) (**c**)

Figure 9. (**a**) YOLOv3 mAP; (**b**) YOLOv4 mAP; (**c**) YOLOv7 mAP.

Some examples of PPE detection through the YOLOv3 [5], YOLOv4 [6] and YOLOv7 [4] models are shown in Figures 10–12. Many different factors are involved in the examples, including individual postures, occlusions, lighting, weather and visuals. In the process of PPE detection, since the target PPE is randomly distributed in the detected image, overlapping occlusion will inevitably occur, and the PPE that is farther from the image acquisition device also occupies a smaller proportion of the image. Consequently, the detection performance of the YOLOv3 [5] model is more prone to false and missed detections. However, the YOLOv4 [6] model proposes a better solution to this problem, as its Mosaic data augmentation randomly combines four images from the training set into one single image on a specific scale. Therefore, the objects in the joined image appear on a smaller scale than the original image, which helps in improving the detection of small objects. Compared with the YOLOv3 [5] model and the YOLOv4 [6] model, the YOLOv7 [4] model greatly improves the detection accuracy in slightly occluded scenes.

Figure 10. Detection examples with YOLOv3 model.

Figure 11. Detection examples with YOLOv4 model.

Figure 12. Detection examples with YOLOv7 model.

4.3. Limitations and Future Work

The current work has several limitations that may provide future directions for improvement. First, most of the images in the dataset collected by the study are based on high brightness and average weather conditions, and although data augmentation effectively addresses this gap, it may still not be representative enough. In the future, we will expand the PPE dataset for more different conditions. Secondly, when the detected object has multiple overlapping objects, the object reaches a certain occlusion ratio, the object size and depth of field are too large and the image of the reflective metal surface or mirror surface may not be able to detect the object correctly. Although there are still some problems with some detection tasks, the situation can be significantly improved by adding more negative samples for training, better data collection and various enhancement methods. Lastly, although the model incorporated four different PPE classes, it was also significant to detect more required PPE in the future, such as safety glasses, industrial gloves, industrial shoes and horizontal lifelines.

5. Conclusions

Deep learning technology has become a powerful solution to learn features directly from data. With the rapid extension of deep learning, the research on PPE detection using CNN has become increasingly practical. While numerous studies in the past have shown that the YOLO series are widely used in areas where accuracy and speed are important, their applicability remains limited in the construction job sites practice, such as privacy issues and the massive computational cost of real-time monitoring.

In this study, a large-scale dataset for detecting PPE compliance was used, including 11,000 pictures of PPE, which can be collected by individuals to provide data support for visual research in the domain of PPE. The proposed model offers a practical detection performance in terms of speed or accuracy. This method has a great application potential in the automatic inspection of PPE components. According to the test results, it can detect with an efficiency of more than 25 FPS and mAP 97%, which can be used to detect whether the construction personnel comply with safety regulations and meet the real-time and high accuracy requirements. However, there is still room for improvement in the proposed model, such as the relatively weak precision and recall of the NO High-visibility vest object.

Author Contributions: Conceptualization, J.-H.L. and L.-K.L.; methodology, J.-H.L.; software, J.-H.L.; validation, J.-H.L., L.-K.L. and C.-C.H.; formal analysis, J.-H.L.; investigation, J.-H.L.; resources, J.-H.L. and C.-C.H.; data curation, J.-H.L. and C.-C.H.; writing—original draft preparation, J.-H.L.; writing—review and editing, L.-K.L.; visualization, J.-H.L.; supervision, L.-K.L.; project administration, L.-K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset link: https://drive.google.com/file/d/1SBlAHoviLHT8uCF0Lhv0ED_Yrdbfwzpq/view?usp=sharing (accessed on 16 December 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ministry of Labor, Taiwan. *Occupational Accident Statistics in 2020*; Ministry of Labor, Taiwan: Taipei City, Taiwan, 2021; pp. 1–2.
2. Occupational Safety and Health Administration. *OSHA Pocket Guide for Construction Safety*; OSHA 3252-05N 2005; Occupational Safety and Health Administration: Washington, DC, USA, 2005; p. 13.
3. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Rose, T.M.; An, W. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* **2018**, *85*, 1–9. [CrossRef]
4. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
5. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
6. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
7. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
8. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
9. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, 7–13 December 2015.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
11. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
12. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *arXiv* **2016**, arXiv:1605.06409.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016.
15. Naticchia, B.; Vaccarini, M.; Carbonari, A. A monitoring system for real-time interference control on large construction sites. *Autom. Constr.* **2013**, *29*, 148–160. [CrossRef]
16. Kelm, A.; Laußat, L.; Meins-Becker, A.; Platz, D.; Khazaee, M.J.; Costin, A.M.; Helmus, M.; Teizer, J. Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites. *Autom. Constr.* **2013**, *36*, 38–52. [CrossRef]
17. Barro-Torres, S.; Fernández-Caramés, T.M.; Pérez-Iglesias, H.J.; Escudero, C.J. Real-time personal protective equipment monitoring system. *Comput. Commun.* **2012**, *36*, 42–50. [CrossRef]
18. Dong, S.; He, Q.; Li, H.; Yin, Q. Automated PPE Misuse Identification and Assessment for Safety Performance Enhancement. *ICCREM* **2015**, 204–214. [CrossRef]
19. Zhang, H.; Yan, X.; Li, H.; Jin, R.; Fu, H. Real-time alarming, monitoring, and locating for non-hard-hat use in construction. *J. Constr. Eng. Manag.* **2019**, *145*, 04019006. [CrossRef]
20. Wu, H.; Zhao, J. An intelligent vision-based approach for helmet identification for work safety. *Comput. Ind.* **2018**, *100*, 267–277. [CrossRef]
21. Xie, Z.; Liu, H.; Li, Z.; He, Y. A convolutional neural network based approach towards real-time hard hat detection. In *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)*; IEEE: Suzhou, China, 2018.
22. Wu, J.; Cai, N.; Chen, W.; Wang, H.; Wang, G. Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. *Autom. Constr.* **2019**, *106*, 102894. [CrossRef]

23. Nath, N.D.; Behzadan, A.H.; Paal, S.G. Deep learning for site safety: Real-time detection of personal protective equipment. *Autom. Constr.* **2020**, *112*, 103085. [CrossRef]
24. Yang, H.; Liu, P.; Hu, Y.; Fu, J. Research on underwater object recognition based on YOLOv3. *Microsyst. Technol.* **2020**, *27*, 1837–1844. [CrossRef]
25. Han, J.; Liao, Y.; Zhang, J.; Wang, S.; Li, S. Target Fusion Detection of LiDAR and Camera Based on the Improved YOLO Algorithm. *Mathematics* **2018**, *6*, 213. [CrossRef]
26. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017.
27. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
28. Zhang, Z.; He, T.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of freebies for training object detection neural networks. *arXiv* **2019**, arXiv:1902.04103.
29. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2014**, *111*, 98–136. [CrossRef]
30. Kosala, R.; Blockeel, H. Web mining research: A survey. *ACM Sigkdd Explor. Newsl.* **2000**, *2*, 1–15. [CrossRef]
31. Tzutalin, D. LabelImg. *GitHub Repos.* **2015**, *6*. Available online: https://github.com/heartexlabs/labelImg (accessed on 1 January 2022).
32. Obaid, H.S.; Dheyab, S.A.; Sabry, S.S. The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*; IEEE: Jaipur, India, 2019.
33. Ilyas, I.F.; Chu, X. *Data Cleaning*; Morgan & Claypool: San Rafael, CA, USA, 2019.
34. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
35. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.