APSIPA Transactions on Signal and Information Processing, 2023, 12, e29

Original Paper

Novel Personal Protective Equipment Detection Technique with Attention-based YOLOv7 and Human Pose Estimation

Krishadawut Olde Monnikhof[1], Punyapat Areerob[1], Zheng Wu[2], Takrit Tanasnitikul[2] and Wuttipong Kumwilaisak[1*]

[1] *Department of Electronics and Telecommunication Engineering, King Mongkut's University of Technology Thonburi, Bangkok, Thailand*
[2] *Panasonic R&D Center, Singapore*

ABSTRACT

This paper proposes a new PPE detection method based on the attention-based YOLOv7 and human pose estimation. The proposed attention module consists of the concatenation of CBAM (Convolutional Block Attention Module) and the SE (Squeeze-and-Excitation) block. This attention module is placed immediately before the detection layer of the YOLOv7 architecture. CBAM derives spatial and channel attention of extracted features from a YOLOv7 backbone. The attention weights prioritize the relevant features in both the spatial and channel domains to be utilized for PPE detection. The SE block refines the attention weights obtained from CBAM before feeding weighted features to the detection layer. Human pose estimation based on YOLO-pose is employed to remove some false positives of PPE detection. The proposed method detects human body parts and assigns key points to human body parts. The essential reference points are computed from the derived vital points. The detection targets far from the reference points will be regarded as

*Corresponding author: Wuttipong Kumwilaisak, wuttipong.kum@kmutt.ac.th. This research was financially supported from King Mongkut's University of Technology Thonburi and Panasonic Singapore Laboratory.

Received 14 March 2023; Revised 03 May 2023
ISSN 2048-7703; DOI 10.1561/116.00000119

false positives and removed. From experimental results, our proposed PPE detection can increase mAP by up to 8.5% at threshold 0.5, 8.8% at threshold 0.5 to 0.95, and reduce false positive detection by 22% on deployment when compared to the original YOLOv7 model.

Keywords: PPE detection, attention-based YOLOv7, human pose estimation, false positive removal.

1 Introduction

Personal Protective Equipment (PPE) is intended to shield workers against illness or harm at work. Generally, PPE includes various equipment such as gloves, vests, helmets, and safety shoes. Based on the statistics from the Occupational Safety and Health Administration, only 16%, 1%, 23%, and 40% of workers who have head injuries, face injuries, foot injuries, and eye injuries wore helmets, face shields, safety shoes, and glasses, respectively. As a result, it is crucial to wear PPE properly since doing so lowers the possibility of being injured in the workplace. Deep learning technology can play a key role in ensuring the proper use of PPE in the workplace. The deep learning model can be deployed to automate PPE detection in the workplace by firing alarms when it detects workers who are not wearing the PPE correctly. In addition, the deep learning model can even predict when PPE needs to be replaced or repaired. This can guarantee that workers are always wearing PPE in good condition.

There have been several research efforts recently to propose automatic PPE detection algorithms. Protik *et al.* [14] deployed YOLOv4 for PPE detection in real-time. They finetuned the already trained YOLOv4 with a public PPE dataset. Ge *et al.* [4] used YOLOX, an anchor-free version of YOLO, to detect PPE. They trained YOLOX with the dataset named CHVG, containing four colored hard hats, vests, and safety glasses. Isailovic *et al.* [6] combined faster R-CNN, MobileNetV2-SSD, and YOLOv7 for PPE detection. Their proposed architecture can exclude false positives well. Nath *et al.* [13] explored three approaches in PPE detection based on YOLOv3. The first approach detected workers, hats, and vests. Then, the model verified whether the individual worker wore the PPE correctly. The second approach detected the workers and simultaneously determined their PPE-wearing compliances by a single CNN model. Finally, the third approach detected only the workers in the input image, which were then cropped and classified by CNN-based classifiers. Gallo *et al.* [3] proposed a real-time detection system based on video streaming and YOLOv4. They evaluated the PPE detection algorithm using the detection performance and interference latency obtained from five CNN

architectures. Wang *et al.* [20] used YOLOv5 for PPE detection. They trained their model on the CHV dataset. The target objects were helmets with four colors, person, and vests. Lo *et al.* [11] deployed multiple versions of YOLO, including YOLOv3, YOLOv4, and YOLOv7, for real-time PPE detection. The models were trained to detect only helmets and vests. Sun *et al.* [17] applied the attention mechanism in YOLOv4 to enhance the detection capability for person detection. Fu *et al.* [2] added the CBAM attention module to YOLOv4 to make the neural network pay more attention to the area that contains more critical information and suppress irrelevant information.

However, the previously mentioned research on PPE detection did not exploit the state-of-the-art YOLO architecture. Most of the current research lacks essential protective gear, such as gloves and shoes, whose intricate designs and small size make them more challenging to detect than hard hats and safety vests. In addition, the attention mechanism that can help the deep learning model focus on the high-potential regions has not been taken into consideration in previous work. As a result, this paper proposes a new PPE detection method based on the attention-based YOLOv7 and human pose estimation. The proposed attention module consists of the concatenation of CBAM (Convolutional Block Attention Module) and the SE (Squeeze-and-Excitation) block. This attention module is placed immediately before the detection layer of the YOLOv7 architecture. CBAM derives spatial and channel attention of extracted features from a YOLOv7 backbone. The attention weights determine the relevant features in both the spatial and channel domains to be utilized for PPE detection. The SE block refines the attention weights obtained from CBAM before feeding weighted features to the detection layer. Human pose estimation based on YOLO-pose is employed to remove some false positives of PPE detection. The proposed method detects human body parts and assigns key points to human body parts. The essential reference points are contributions of this paper can be summarized as follows.

1. By adding an attention mechanism to YOLOv7, we improve the ability to focus on the most crucial regions of features by selectively weighting different regions of the features before detection.

2. We combine CBAM and SE attention modules to improve the performance of PPE detection. After the CBAM first extracts global context information from the input data, the SE blocks perform channel-wise re-weighting on the attended feature maps.

3. We reduce the false positive detection using human pose estimation by mapping the detected PPE to its appropriate position on various human body parts, then deleting the detected PPE from locations where its position does not correspond to the regular wearing position.

This paper is organized as follows. Section 2 describes the novel enhanced YOLOv7 architecture with an attention mechanism. The overall architecture of YOLOv7 is presented in Section 2.1. The proposed combination of CBAM and the SE block to prioritize extracted features is described in Section 2.2. Section 2.3 discusses the false positive removal based on human pose estimation. Experimental results are provided in Section 3. Finally, concluding remarks are in Section 4.

2 Personal Protective Equipment (PPE) Wearing Detection Method with YOLOv7 and Attention Mechanism

2.1 YOLOv7

The YOLO (You Only Look Once) v7 model [19] is one of the latest object detection frameworks in the family of YOLO models. YOLOv7 can achieve more accurate bounding box prediction than its predecessors while maintaining exact inference times, gratitude to its new extended efficient layer aggregation, model scaling, re-parameterization planning, and auxiliary head coarse-to-fine. The main architecture of YOLOv7 is still the same as its predecessors as consisting of the backbone, neck, and head as illustrated in Figure 1.

The backbone of YOLOv7 extracts essential features from an input image and then feeds them to the head. In general, the backbone architecture of previous YOLO versions is based on convolutional neural networks such as Darknet-53 (YOLOv3 [15] and CSPDarnet-53 (YOLOv4, YOLOv5 [1, 7]). The Extended Efficient Layer Aggregation Network (E-ELAN) is introduced to the backbone of YOLOv7 to help the model learn better. Furthermore, the YOLOv7 backbone can be scaled to match different levels of required accuracy and inference speeds. It can be done by adjusting the network depth, width, and layer concatenation. YOLOv7 adopts module-level re-parameterization techniques to improve inference performance by averaging model outputs. To be more specific, during training, the training session is split into multiple training modules. Then, the weights of each module are ensembled to obtain the final model. For example, the E-ELAN computation block's 3×3 convolutional layers can be replaced by the combination of RepConv [19], identity connection, and convolutional layer. The outputs of different E-ELAN computation blocks are finally averaged.

The neck of YOLO mainly relies on the variant feature pyramid network introduced in YOLOv3 [15]. The Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PANet) are utilized in YOLOv4 [1]. The PANet in YOLO adds a path aggregation module to the network architecture, which helps improve object detection accuracy by aggregating features from multiple scales and resolutions. This module effectively handles objects of various sizes and shapes.

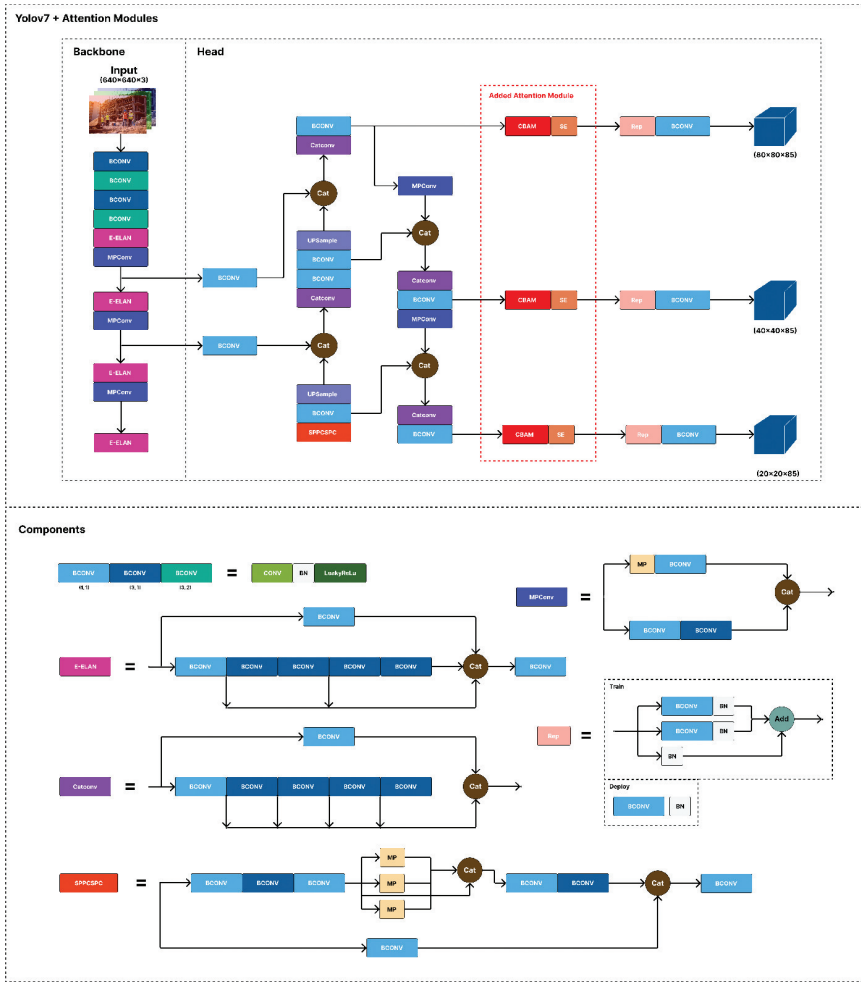

Figure 1: Enhanced YOLOv7 architecture with attention mechanism.

Finally, the head consists of output layers, which provide the final object detection. In YOLOv7, the head creates feature pyramids by collecting feature maps extracted from the backbone and outputting the final predictions. The head part in YOLOv7 uses the SPPCSP (Spatial Pyramid Pooling Cross Stage Partial) that combines the SPP (Spatial Pyramid Pooling) neck and the CSP (Cross Stage Partial) backbone and ELAN modules to aggregate image features. RepConv then modifies the output feature channels. Finally, the YOLOv7 output and prediction are derived from the 1×1 convolution.

2.2 PPE Wearing Detection with Enhanced YOLOv7 based on Attention Mechanism

We integrate the attention mechanism into the YOLOv7 architecture in this section to improve its detection of PPE. The attention mechanism improves the model's ability to automatically focus on the most relevant PPE regions from the input images by weighing image regions differently. In this paper, we utilize the combination of CBAM (Convolutional Block Attention Module) [21] and the SE Block (Squeeze-and-Excitation Block) [5] as our attention mechanism. CBAM combines channel attention with spatial attention. Spatial attention weighs the significance of each position in the spatial dimensions of the feature maps, whereas channel attention weighs the significance of each channel in the feature maps. The SE Block gives the feature maps more attention by learning about each channel in the feature maps through a Squeeze-and-Excitation process. We will first describe the architecture of CBAM and SE Block in detail in the following sections.

2.2.1 Convolutional Block Attention Module (CBAM)

Suppose the PPE feature map from YOLOv7 is $\mathbf{F}$ with dimensions $C \times H \times W$. CBAM generates channel and spatial attention maps, as shown in Figure 2.

The channel attention module applies average pooling and max pooling to $\mathbf{F}$ as

$$\mathbf{F}_{avg} = AvgPool_s(\mathbf{F}), \tag{1}$$

$$\mathbf{F}_{max} = MaxPool_s(\mathbf{F}), \tag{2}$$

and then processes the pooled feature maps with multi-layer perceptron (MLP) and sigmoid activation as

$$M_c(\mathbf{F}) = \sigma(MLP(\mathbf{F}_{avg}) + MLP(\mathbf{F}_{max})), \tag{3}$$

$$MLP(\mathbf{F}_{avg}) = W_1(ReLU(W_0(\mathbf{F}_{avg}))), \tag{4}$$

$$MLP(\mathbf{F}_{max}) = W_1(ReLU(W_0(\mathbf{F}_{max}))), \tag{5}$$

where $M_c(\mathbf{F})$ is the channel attention map with dimensions $C \times 1 \times 1$, and r is the reduction ratio.

The spatial attention module applies global average pooling and max pooling along the channel dimension as

$$\mathbf{F}_{savg} = AvgPool_c(\mathbf{F}), \tag{6}$$

$$\mathbf{F}_{smax} = MaxPool_c(\mathbf{F}), \tag{7}$$

followed by concatenation and convolutional operation, resulting in

$$M_s(\mathbf{F}) = \sigma(Conv_{7\times7}([\mathbf{F}savg; \mathbf{F}smax])), \tag{8}$$

where $M_s(\mathbf{F})$ is the spatial attention map with dimensions $1 \times H \times W$.

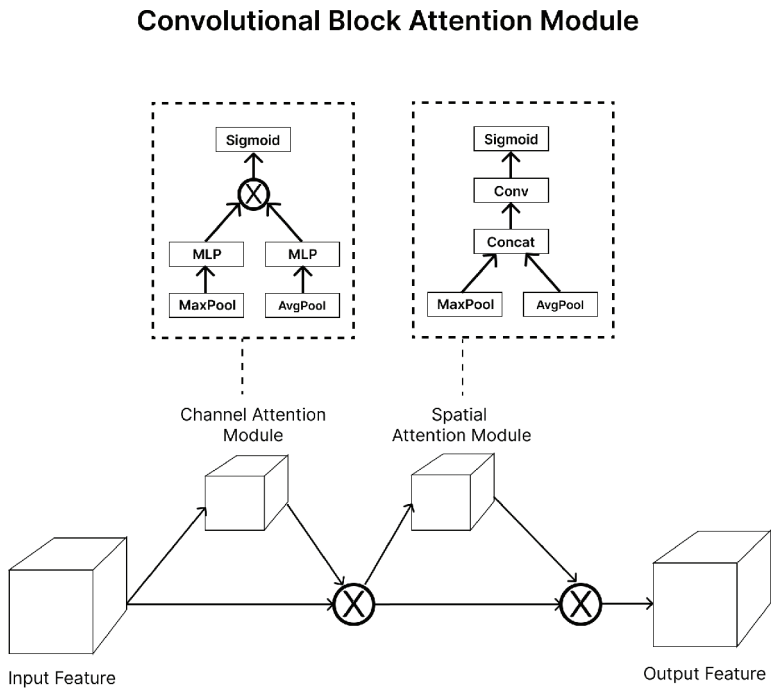

Figure 2: CBAM architecture.

2.2.2 Squeeze-Excitation (SE) Block

The SE block assigns channel-specific weights to a feature map, with three operations: squeeze, computation, and excitation, as shown in Figure 3.

The squeeze operation applies global average pooling to an input feature $\mathbf{F}$ with dimensions of $C \times H \times W$, resulting in a $C \times 1 \times 1$ pooled feature $\mathbf{F}_{squeeze}$

$$\mathbf{F}_{squeeze} = AvgPool(\mathbf{F}). \tag{9}$$

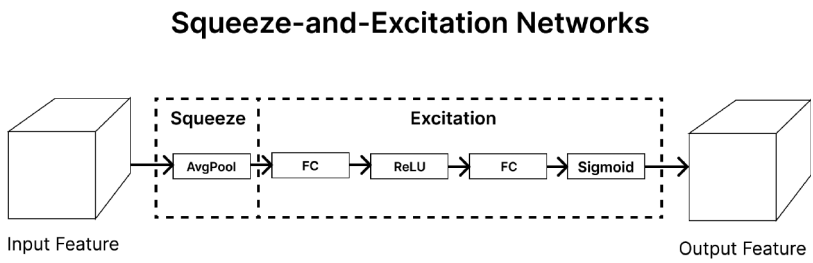

Figure 3: SE block architecture.

This feature vector passes through two fully connected layers with ReLU and sigmoid activation functions, respectively, resulting in a $C \times 1 \times 1$ feature $\mathbf{F}_{operation}$

$$\mathbf{F}_{operation} = \sigma(W^{se}1(ReLU(W^{se}0(\mathbf{F}_{squeeze})))). \tag{10}$$

Finally, the excitation operation uses $\mathbf{F}_{operation}$ as a per-channel weight vector, generating the SE attention map $M_{se}(\mathbf{F})$ with dimensions $C \times H \times W$.

2.2.3 CBAM-SE Attention Mechanism

We combine CBAM and the SE block to be CBAM-SE attention model. The extracted features of YOLOv7 are fed to CBAM to compute the attention maps in both spatial and channel dimensions. The derived attention maps are applied to the original features to obtain the prioritized feature map as

$$\mathbf{F}_c = M_c(\mathbf{F}) \otimes \mathbf{F}, \tag{11}$$

$$\mathbf{F}_{cbam} = M_s(\mathbf{F}_c) \otimes \mathbf{F}_c, \tag{12}$$

where $\mathbf{F}_{cbam}$ is the prioritized PPE feature map from CBAM. Then, $\mathbf{F}_{cbam}$ is passed to the SE Block to reassign the priorities of CBAM channel features. This process is responsible by the excitation operation of the SE block, which is

$$\mathbf{F}_{cbam-se} = M_{se}(\mathbf{F}_{cbam}) \otimes \mathbf{F}_{cbam}, \tag{13}$$

where $\mathbf{F}_{cbam-se}$ is the output feature map from the CBAM-SE attention module.

Placement positions of attention modules within YOLOv7 affect the overall PPE detection performance. For instance, if we place the attention model immediately after the input images, the final PPE detection results will appear suboptimal. This is based on the fact that PPE locations can be anywhere within input images. As a result, the attention model is unable to properly prioritize regions corresponding to PPE. To solve this problem, we place the proposed attention module immediately before the detection layer, as shown in Figure 4. At the underlying location, the attention model can only prioritize essential PPE features extracted from the YOLOv7 backbone and neck, which are generally less location-dependent.

2.3 PPE False Detection Removal using Human Pose Estimation

This section proposes a method to improve the PPE detection results obtained from our attention-based YOLOv7. Because most PPE-related images and videos involve people wearing PPE suits, it is reasonable to use human bodies as key indications to remove false positives from the PPE detection results. The simplest method is first to detect humans and then to compute the

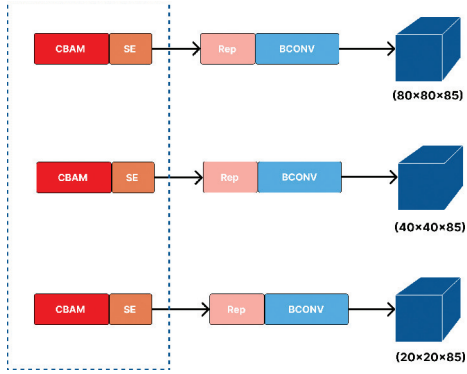

Figure 4: The position of the CBAM-SE attention module in YOLOv7.

Intersection over Union (IOU) between the detected PPE objects and humans. If these PPE objects are within the human bounding boxes, we can assume that the detected PPE results should not be false positives. Unfortunately, when multiple people are within the same image, human bounding boxes may cover almost the whole image region. False positives likely fall inside these bounding boxes and will not be eliminated.

To solve this problem, we identify not only human locations but also key locations corresponding to different body parts. With derived human key points and detected PPE, we map the detected targets with human body parts corresponding to key points. Specifically, we deploy YOLO-pose [12] to input images simultaneously with our enhanced YOLOv7. The detected objects not corresponding to the human key points are discarded as false positives. Figure 5 shows the human key points obtained from YOLO-pose.

Figure 5: Key points on a human body obtained from YOLO-pose.

Targeted wearing PPE equipment consists of four classes: helmet; vest; gloves; and shoes. We can directly relate the detected object to key points corresponding to the wrists and ankles for gloves and shoes. However, there are multiple key points agreeing with the human head and the human body. As a result, we need to find the reference points of the head and the body by averaging the key point coordinates. Suppose that coordinate sets of the human body and head key points are

$$\Omega = \{(x_{b,1}, y_{b,1}), (x_{b,2}, y_{b,2}), \ldots, (x_{b,m}, y_{b,m})\}, \tag{14}$$

$$\Theta = \{(x_{h,1}, y_{h,1}), (x_{h,2}, y_{h,2}), \ldots, (x_{h,n}, y_{n,m})\}, \tag{15}$$

where m and n are key numbers of the body and the head,respectively. Then, the reference points of the body and the head can be expressed as

$$x_b = \frac{\sum_{i=1}^{m} x_{b,i}}{m}, \tag{16}$$

$$y_b = \frac{\sum_{i=1}^{m} y_{b,i}}{m}, \tag{17}$$

$$x_h = \frac{\sum_{i=1}^{n} x_{h,i}}{n}, \tag{18}$$

$$y_h = \frac{\sum_{i=1}^{n} y_{h,i}}{n}, \tag{19}$$

where (x_b, y_b) and (x_h, y_h) are reference points of the body and the head, respectively.

We use the distance between the reference point to the center of the bounding box to determine the relevance of the detected object and the key point. If the center coordinate of the bounding box is (x_c, y_c), the distances between reference points to the bounding box can be calculated via

$$D_{bc} = \sqrt{(x_b - x_c)^2 + (y_b - y_c)^2}, \tag{20}$$

$$D_{hc} = \sqrt{(x_h - x_c)^2 + (y_h - y_c)^2}, \tag{21}$$

where D_{bc} and D_{hc} are the distances from the body reference point and the head reference point to the center of the bounding box separately. We compare D_{bc} and D_{hc} with the threshold. The bounding box is declared a false positive if the distance is greater than a threshold. Since the values of D_{bc} and D_{hc} are affected by the size of the bounding box, it is reasonable to take the bounding box size into consideration in order to determine the threshold. In other words, the smaller size the human appearance in the image, the comparatively smaller the values of D_{bc} and D_{hc}.

For the bounding box with the size of $w_d \times h_d$, the threshold used to decide whether it is a false positive is reckoned via

$$T_{adj} = \sqrt{(w_d^2 + h_d^2) \times T}, \tag{22}$$

where T is a hyperparameter related to a threshold. If $D_{bc} > T_{adj}$, then this bounding block is a false positive of a vest class. If $D_{hc} > T_{adj}$, this bounding block is considered as a false positive of a helmet class. Moreover, we can also further assume that one person normally wears only one helmet, one vest, two gloves, and one pair of shoes. Hence, we only maintain these sets of detected PPE with smallest distances to the human key points. Other surplus detected objects will be considered as false positives and then discarded. The steps of false positive removal is visualized in Figure 6.

Figure 6: False Positive Removal Mechanism.

3 Experimental Result

3.1 Evaluation Metrics

This section evaluates the PPE-wearing detection performances of the proposed attention-based YOLOv7 and false positive removal algorithm. The assessment metric is the mean Average Precision (mAP). To compute mAP, we must deploy the Confusion Matrix, Intersection over Union (IoU), Precision, and Recall. The Confusion Matrix contains four attributes: True Positives (TP); True Negatives (TN); False Positives (FP); and False Negatives (FN). True positives and false positives are caused by the model detecting the PPE objects, and the objects are a part of the ground truth and not a part of the ground truth, respectively. True positives and false positives occur when the model detects PPE objects that are, or are not, part of the ground truth, respectively. Conversely, true negatives and false negatives arise when the model does not detect PPE objects that are not, or are, part of the ground truth, respectively. The IoU is defined as the ratio of the intersection area between the predicted bounding box and the ground truth over the union area of the predicted bounding box and the ground truth. Based on the previously described metrics, the Precision can be computed as

$$Precision = \frac{n_{TP}}{n_{TP} + n_{FP}}, \tag{23}$$

where n_{TP} and n_{FP} are the numbers of true positives and false positives, respectively. Moreover, the Recall can be expressed as

$$Recall = \frac{n_{TP}}{n_{TP} + n_{FN}}, \tag{24}$$

where n_{TP} and n_{FN} are the numbers of true positives and false negatives, respectively.

Average Precision (AP) is calculated by averaging Precision results from different thresholds. The threshold is dictated by the IoU values. For example, suppose that we set the IoU threshold as 0.5. If two detected PPE possess 0.3 and 0.5 values of IoU, we can declare that these two objects are a false positive and a true positive, respectively. Since in our application, there are four classes of PPE equipment: helmet; vest;gloves; and shoes. The mean Average Precision (mAP) can be computed from averaging the AP values of these four classes, which is

$$mAP = \frac{1}{4}(AP_h + AP_v + AP_g + AP_s), \tag{25}$$

where AP_h, AP_v, AP_g, and AP_s are Average Precision of class helmet, vest, gloves, and shoes, respectively.

3.2 Dataset and Data Preprocessing

The PPE dataset used for train our model contains 1500 images and 5,286 annotations. The data distributions among train, validation, and test sets of PPE equipments are shown in Figure 7. The example of annotated image can be illustrated in Figure 8. We exploit the data augmentation techniques including image rotation, color distortion, image translation, and image transformation to increase the training data. As a result, the over fitting can be avoided.

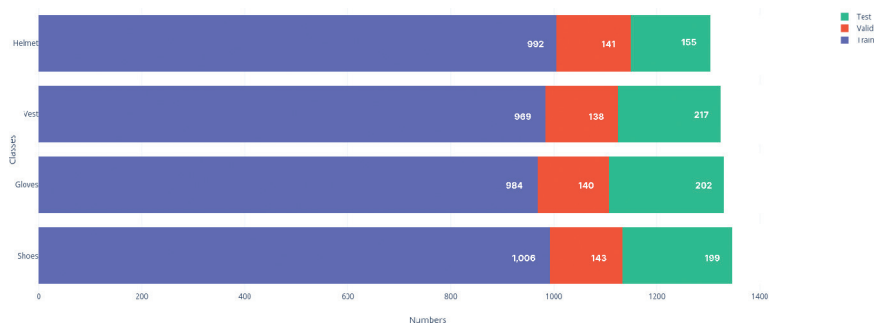

Figure 7: The Annotation Numbers of Train, Validation, and Test Sets for PPE Detection.

3.3 Performance Evaluation of Attention-based YOLOv7

This section gives the evaluated performance of the proposed attention-based YOLOv7 over the existing methods. Since PPE-wearing detection methods are generally deployed in real-time, we benchmark the proposed method with

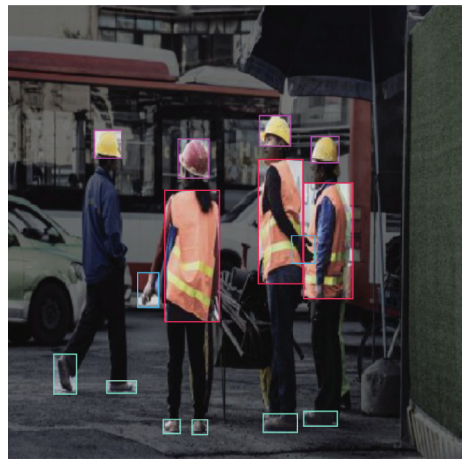

Figure 8: Example of Annotated Image.

YOLOv5 [7], YOLOv7 [19], SSD [10], EfficientDet [18], RetinaNet [8], and Faster-RCNN [16]. The hyperparameters of these models are set to their default values. We train all models on the same dataset using NVIDIA GeForce RTX 3090 10-GB GPUs, AMD Ryzen Threadripper 2950X 16-Core Processor, 3.500 GHz CPU, and 130 GB RAM. Table 1 compares the PPE detection performance of YOLOv7 with other detection modules by employing mAP and inference time. The result in terms of mAP@0.5 of YOLOv7 and EfficientDet is very similar. However, YOLOv7 outperforms other modules in terms mAP@0.5:0.95, a lot while still having a relatively high inference time regarding frames per second.

As we previously discussed in Section 2.2.3, the placement position of the attention module plays an essential role in improving PPE-wearing detection accuracy. As a result, we conduct experiments by placing the SE attention

Table 1: Performance comparison of PPE wearing detection of different detection models.

Architecture	Backbone	mAP@.5	mAP@.5:.95	FPS
YOLOv5	PANet	0.763	0.419	**96**
YOLOv7	E-ELAN	0.811	**0.515**	84
SSD	VGG16	0.701	0.310	44
EfficientDet	EfficientDet-D0	**0.813**	0.472	41
RetinaNet	ResNet50	0.742	0.354	27
RetinaNet	ResNet101	0.764	0.372	20
Faster-RCNN	ResNet50	0.768	0.376	26
Faster-RCNN	ResNet101	0.811	0.389	20

Table 2: PPE-wearing performance regarding placement positions of the attention module.

Attention Module Position	P	R	mAP@.5	mAP@.5:.95
Placing after the input	0.891	0.711	0.796	0.481
Placing after the head	0.872	0.723	0.808	0.492
Placing before the detection layer	**0.911**	**0.768**	**0.831**	**0.534**

module in a different position within the YOLOv7 architecture. We intentionally deploy only the SE block in this testing because the SE block requires less computation complexity. The results of this study can be presented in Table 2. We observe for mAP at IoU threshold 0.5 placing the attention module after the input performs worst at 0.796, whereas placing the attention module before the detection layer provides the best mAP value. With the deployment of mAP at IoU between 0.5 and 0.95 (i.e., if the IoU value is between this interval, the PPE detection is declared), the results are consistent with the mAP value 0.5.

The attention mechanism can best focus on the most important features and suppress those that are irrelevant when the attention module is placed before the detection layer. The attention module may not be as effective if it comes after the input because the input features may still be too raw or coarse to properly capture the relevant information. Although the results are still lower than before the detection layer, placing the attention module after the head produces better results than after the input. This implies that the features acquired at this level may not be as discriminating as those that were accessible right before the detection layer.

As a result, we decided to position our proposed attention module before the detection layer.

Tables 3 and 4 show the results of enhanced YOLOv7 with different combinations of CBAM and the SE block. The results are obtained from the training with batch size, learning rate, and training epoch equaling 16, 0.01, and 200, respectively. The enhanced YOLOv7 corresponding to Table 4 deploys the pre-trained weights from the COCO 2017 dataset [9], whereas the model in Table 3 does not. We found that pre-trained weights from the COCO 2017 dataset gave inferior results to the model explicitly trained from the PPE dataset. To be more specific, the improvement is up to 8.5% for the mAP with a threshold of 0.5 and 8.8% with a threshold between 0.5 and 0.95. The combination of YOLOv7 with SE and CBAM in parallel improves performance, although not as much as when both attention mechanisms are coupled sequentially. This further supports the hypothesis that sequential integration of attention mechanisms allows each mechanism to refine the features successively, leading to better overall performance.

Figure 9 shows the confusion matrix of YOLOv7 + CBAM + SE (before false positive removal). The model correctly identified the presence of each

Table 3: PPE-wearing detection performance of the enhanced YOLOv7 with different combinations of CBAM and the SE block, when no pretrained weights are used.

Architecture	P	R	mAP@.5	mAP@.5:.95
YOLOv7	0.877	0.739	0.811	0.515
YOLOv7 + CBAM	0.876	0.745	0.822	0.524
YOLOv7 + SE	0.894	0.743	0.819	0.523
YOLOv7 + CBAM + SE	**0.911**	**0.768**	**0.831**	**0.534**
YOLOv7 + SE + CBAM	0.886	0.751	0.822	0.519
YOLOv7 + SE & CBAM in parallel	0.896	0.728	0.809	0.523

Table 4: PPE-wearing detection performance of the enhanced YOLOv7 with different combinations of CBAM and the SE block, when the pretrained weights from COCO 2017 dataset are used.

Architecture	P	R	mAP@.5	mAP@.5:.95
YOLOv7	0.878	0.744	0.814	0.534
YOLOv7 + CBAM	0.894	0.857	0.884	0.599
YOLOv7 + SE	**0.919**	0.836	0.888	0.609
YOLOv7 + CBAM + SE	0.916	**0.863**	**0.899**	**0.622**
YOLOv7 + SE + CBAM	0.899	0.834	0.886	0.612
YOLOv7 + SE & CBAM in parallel	0.934	0.829	0.891	0.605

PPE category as follows: 62% for gloves, 87% for helmets, 72% for shoes, and 85% for vests.

3.4 Performance Evaluation of False Positive Removal with Human Pose Estimation

This section evaluates the improved performance of PPE-wearing detection with the proposed false positive removal method. We apply the false positive removal engine to the detection results of enhanced YOLOv7 with attention mechanism. The results are shown in Table 5.

Figure 9 shows the confusion matrix before and after applying the false positive removal to the PPE detection using YOLOv7 + CBAM + SE. The redundant false positive detection of gloves and shoes, whose intricate designs and small size make them more confused and lead to a lot of false positives, are significantly decreased. Moreover, it makes the model more precise and decreases the false positive detection of helmet and gloves too.

When integrating the false positive removal on the deployment to the enhanced YOLOv7, we can reduce 22% of the false positive detection after

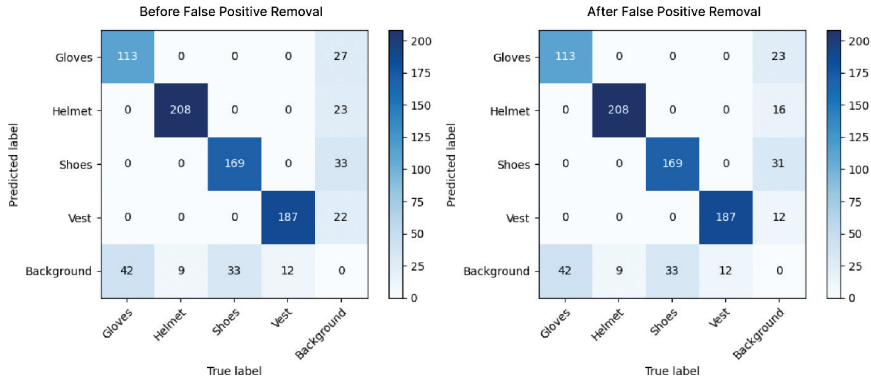

Figure 9: Confusion Matrices Before and After Applying the False Positive Removal.

Table 5: PPE-wearing detection performance (total true positive and total false positive) when applying false positive removal in different environments.

Env.	Conf. Treshold	IoU Treshold	FP Removal	Total TP	Total FP
Testing	0.001	0.65	No	742	6492
Deployment	0.25	0.45	No	677	105
Testing	0.001	0.65	Yes	726	1586
Deployment	0.25	0.45	Yes	677	82

applying the enhanced YOLOv7 with attention mechanism without sacrificing any true positive detections.

Specifically, the false positive detection rates decreased by 15% for gloves, 30% for helmets, 6% for shoes, and 45% for vests, demonstrating the effectiveness of the false positive removal algorithm in enhancing the detection performance for each PPE category.

The comparison images for PPE detection from the original YOLOv7 model, the YOLOv7 model with our attention module added (CBAM + SE), and our model after employing false positive removal to the attention-based YOLOv7 are shown in Figure 10.

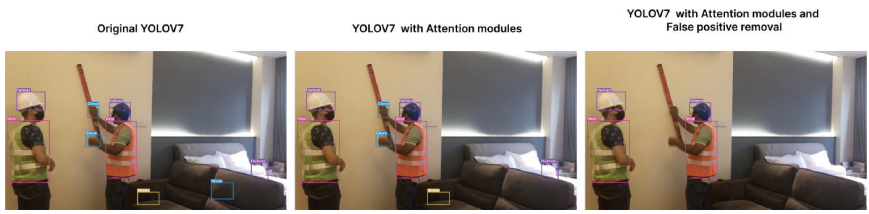

Figure 10: Result Comparison Images for PPE Detection.

4 Conclusions

In this research, we introduced a new method for PPE detection by adding an attention module to YOLOv7. The attention model was based on the concatenation of the Convolutional Block Attention Module (CBAM) and Squeeze-Excitation (SE) block. The CBAM computed attention maps across spatial and channel dimensions. The SE block refined the attention maps obtained from the CBAM. This attention module was placed before the detection layer of YOLOv7. The false positive removal algorithm was also proposed based on the human pose estimation to improve the detection accuracy. Our results improved mAP up to 8.5% at threshold 0.5, 8.8% at threshold 0.5 to 0.95, and a 22% reduction in false positive detection compared to the original YOLOv7 model. With this PPE detection model, it is possible to gain real-time monitoring accuracy and enforce compliance with PPE to establish employee safety.

References

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020, https://arxiv.org/abs/2004.10934.

[2] H. Fu, G. Song, and Y. Wang, "Improved YOLOv4 Marine Target Detection Combined with CBAM," *Symmetry*, 13, 2021, DOI: 10.3390/sym13040623.

[3] G. Gallo, F. D. Rienzo, F. Garzelli, P. Ducange, and C. Vallati, "A Smart System for Personal Protective Equipment Detection in Industrial Environments Based on Deep Learning at the Edge," *IEEE Access*, 10, 2022, 110862–78, DOI: 10.1109/ACCESS.2022.3215148.

[4] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," 2021, https://arxiv.org/abs/2107.08430.

[5] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," 2017, https://arxiv.org/abs/1709.01507.

[6] V. Isailovic, A. Peulic, M. Djapan, M. Savković, and A. Vukicevic, "The compliance of head-mounted industrial PPE by using deep learning object detectors," *Scientific Reports*, 12, 2022, DOI: 10.1038/s41598-022-20282-9.

[7] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, TaoXie, J. Fang, Lorna, Z. Yifu, C. Wong, V. Abhiram, D. Montes, Z. Wang, C. Fati, J. Nadar, V. Sonck, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain, *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation*, version v7.0, November 2022, https://doi.org/10.5281/zenodo.7347926.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," 2017, https://arxiv.org/abs/1708.02002.

[9] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," 2014, https://arxiv.org/abs/1405.0312.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, Springer International Publishing, 2016, 21–37, https://doi.org/10.1007%2F978-3-319-46448-0_2.

[11] J.-H. Lo, L.-K. Lin, and C.-C. Hung, "Real-Time Personal Protective Equipment Compliance Detection Based on Deep Learning Algorithm," *Sustainability*, 15, 2022, 391, DOI: 10.3390/su15010391.

[12] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss," 2022, https://arxiv.org/abs/2204.06806.

[13] N. Nath, A. Behzadan, and S. Paal, "Deep learning for site safety: Real-time detection of personal protective equipment," *Automation in Construction*, 112, 2020, 103085, DOI: 10.1016/j.autcon.2020.103085.

[14] A. Protik, A. Hossain, and S. Siddique, "Real-time Personal Protective Equipment (PPE) Detection Using YOLOv4 and TensorFlow," 2021, DOI: 10.1109/TENSYMP52854.2021.9550808.

[15] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018, https://arxiv.org/abs/1804.02767.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," 2015, https://arxiv.org/abs/1506.01497.

[17] J. Sun, H. Ge, and Z. Zhang, "AS-YOLO: An Improved YOLOv4 based on Attention Mechanism and SqueezeNet for Person Detection," 2021, 1451–6, DOI: 10.1109/IAEAC50856.2021.9390855.

[18] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," 2019, https://arxiv.org/abs/1911.09070.

[19] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, https://arxiv.org/abs/2207.02696.

[20] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, and Y. Zhao, "Fast Personal Protective Equipment Detection for Real Construction Sites Using Deep Learning Approaches," *Sensors*, 21, 2021, 3478, DOI: 10.3390/s21103478.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," 2018, https://arxiv.org/abs/1807.06521.