Journal of
Information Systems Engineering
and Business Intelligence

Vol.11, No.3, October 2025
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

CBTi-YOLOv5: Improved YOLOv5 with CBAM, Transformer, and BiFPN for Real-Time Safety Helmet Detection

Tio Dharmawan [1)*], Danang Setiawan [2)], Muhamad Arief Hidayat [3)], Vandha Pradwiyasma Widartha [4)]

[1)2)3)]*Faculty of Computer Science, Universitas Jember, Jember, Indonesia*
[1)]tio.pssi@unej.ac.id, [2)]danangst113@gmail.com, [3)]arief.hidayat@unej.ac.id

[4)]*Department of Information System, Pukyong National University, Busan, South Korea*
[2)]vandhapw@pukyong.ac.kr

Abstract

Background: Some construction workers are often in a situation where injuries can occur from negligence in the use of safety helmets. To avoid this, supervision of the use of safety helmets should be conducted continuously during the work process through the application of computer vision technology. However, the complex background of the construction environment is a challenge to detecting small and densely packed safety helmets accurately.
Objective: The construction environment is complex, and the wide workspace allows workers to be in an area far from supervision. The process makes it difficult for models to detect the use of safety helmets in complex, wide, and very high object density construction environments. Therefore, this study aims to overcome the problem by modifying YOLOv5s (You Only Look Once version 5) architecture.
Methods: Real-time monitoring of the use of safety helmets could be performed using YOLOv5. This study proposed a modified YOLOv5s model called CBTi-YOLOv5s. The model incorporated Convolutional Block Attention Module (CBAM), Transformer, and Bi-directional Feature Pyramid Network (BiFPN) to improve feature extraction, multi-scale object representation, as well as detection accuracy, specifically on small and high-density objects in complex construction environments.
Results: The results showed the modified YOLOv5s architecture had made an improvement of 3.7% in mean average precision (mAP) compared to the base YOLOv5s model. mAP of the base YOLOv5s model was 93.6%, while the modified CBTi-YOLOv5s model achieved 97.3%. The proposed modified YOLOv5s model also achieved an inference speed of 58 frames per second (FPS), and the base model achieved 104 FPS.
Conclusion: CBTi-YOLOv5s improved the accuracy, mAP, and ability to detect objects of varying scales. However, this improvement had drawbacks, namely increased model size and decreased inferential speed due to increased model architectural complexity..

Keywords: Bi-FPN, CBAM, CBTi-YOLOv5s, Helmet Detection, Transformer, YOLOv5

Article history: Received 14 August 2024, first decision 18 April 2025, accepted 14 August 2025, available online 28 October 2025

I. INTRODUCTION

Occupational safety and health (OHS) is crucial in forming safe working environments and is essential in preventing both physical as well as mental health issues on the job [1]. A major aspect of OHS is ensuring the proper use of personal protective equipment (PPE), specifically safety helmets, which are crucial in protecting workers from head injuries caused by falling or flying objects. In Indonesia, it is reported that around 60% of workplace head injuries happen because workers are not wearing safety helmets [2], showing the need for better safety measures. A promising solution to this problem is vision-based detection systems, which can automatically monitor helmet use and help prevent accidents as well as fatalities at work

Among the various methods for monitoring PPE, vision-based methods are often favored because the tools are flexible and non-intrusive, different from sensor-based systems that can be more burdensome [3]. To be effective,

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.11.3.353-363

these systems need to detect objects quickly and accurately in real-time, specifically in the complex, ever-changing environments typical of construction sites.

Earlier solutions have applied a combination of deep learning and traditional machine learning, such as using FaceNet for feature extraction and Support Vector Machines (SVM) for PPE detection [4]. As these methods can be quite accurate, the tools tend to comprise multiple stages, which slows down processing and makes the models less ideal for real-time use. Relating to this discussion, fully deep learning-based models have surfaced to overcome this limitation. An example is PerspectiveNet, which leverages EfficientNet v2 as its backbone and is optimized for use on embedded systems [5]. However, these models can struggle in cluttered scenes or when trying to detect small objects [6], which are common challenges in real-world construction settings.

Single-stage object detection models, particularly You Only Look Once (YOLO) family, have advantages including fast inference speeds and solid accuracy. For instance, an improved version of YOLOv4 (YOLO version 4) has reached a mean average precision (mAP) of 91.03% for detecting safety helmets in aerial images [7]. To handle the limited computing power, YOLO-PL is developed. YOLO-PL is a lightweight variant of YOLOv4, which improved the detection speed on constrained resource systems [8].

Numerous models are developed with an inherent drawback despite the strengths, prioritizing speed or accuracy. Consequently, the models often underperform in real-world applications consisting of small, overlapping, or densely clustered objects, particularly in complex environments such as construction sites.

YOLO architecture has tremendous advancements, and YOLOv5 (YOLO version 5) marked a turning point, offering several model variants modified to different devices as well as performance needs [9], [10], [11], [12], [13], [14]. Irrespective of its widespread use, YOLOv5 base models still have some significant drawbacks, including struggles with multi-scale object detection, specifically in crowded or complex scenes. The lightweight versions advance the computational speed but sacrifice feature richness, which can affect detection accuracy [13], [14].

The studies have relied on base YOLOv5 models for safety helmet detection tasks [3], without any architectural improvements to tackle these limitations. Similarly, earlier models such as YOLOv4 and YOLO-PL lack features including attention mechanisms or advanced multi-scale feature fusion, both of which are increasingly recognized for improving detection robustness in real-world applications.

This study introduces an improved version of YOLOv5s by incorporating three major architectural modules to address the limitations. First, Convolutional Block Attention Module (CBAM) is incorporated to refine spatial and channel-wise feature representation. Second, a Transformer encoder is added to improve global feature extraction and positional awareness to handle dense object scenarios. Third, Bi-directional Feature Pyramid Network (BiFPN) is used to strengthen multi-scale feature fusion. These components aim to produce a lightweight, accurate, and real-time object detection model, optimized for monitoring safety helmet usage in construction environments.

The structure used during the process of this study is as follows. Section II reviews related work and background studies, Section III describes the proposed model architecture, and Section IV explains the experimental setup and presents evaluation results. In addition, Section V discusses the results, and Section VI concludes the study.

II. Literature Review

Helmet is a crucial part of PPE for construction workers, providing essential protection against potentially life-threatening head injuries. However, many workers frequently fail to wear helmets consistently, which significantly increases the risk. Object detection technologies have surfaced as a promising solution for automatically monitoring helmet use, enabling real-time surveillance to help create safer work environments. Deploying the systems is challenging as the implementation in construction sites requires both accuracy and the efficient computational performance of detection methods, making the models practical for real-world, real-time applications.

A hybrid method for detecting safety helmets in video footage was introduced by combining machine learning and image processing methods, comprising three stages, namely face detection using Haar-like features, motion filtering, and hard hat color detection [15]. This method struggled with distinguishing false positives from actual faces, although the inclusion of color information provided some filtering capability. A CCTV-based monitoring system was developed to detect faces using Haar-like features and identify helmets based on red color as well as shape outline criteria, where the system activated warnings when workers were detected without helmets [16]. As traditional machine learning methods often rely on handcrafted features and rule-based detection have shown moderate success in helmet detection, the advent of deep learning has substantially improved accuracy and robustness through automatic feature extraction from complex visual data.

Recent advances in deep learning have significantly influenced object detection. Currently, the most effective object detection algorithms fall into two main categories, namely multi-stage and single-stage detectors [17]. Multi-stage detectors, such as Faster R-CNN, offer superior detection accuracy but at the cost of high computational

complexity, making the systems less suitable for real-time applications. Single-stage detectors, including YOLO, prioritize faster inference speeds [6] with relatively lower computational requirements, often at a slight expense in accuracy [18]. Among these, YOLO family of models has acquired substantial attention for the impressive speed-accuracy drawback.

Several studies have applied YOLO models to the task of safety helmet detection. For instance, [19] evaluated various YOLOv5 versions YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, achieving mAP scores of 94.2%, 95.3%, 95.5%, 95.6%, as well as 95.8%, respectively. Although these results show the effectiveness of YOLOv5, baseline models still present limitations when detecting small, dense, and multi-scale objects under real-world conditions. Studies by [20][21] proposed improving detection performance through the incorporation of CBAM and BiFPN to address object scale variation challenges. The method used during the analysis achieved an improvement of 1.6% in mAP and 5.3% in precision, significantly improving multi-scale object recognition and feature representation. In another study [20], Transformer module was introduced alongside BiFPN and CBAM to further improve the feature extraction capabilities, particularly for dense object scenarios. This method led to a 5.7% increase in precision and a 1.6% improvement in mAP compared to the baseline model. Similarly, [22] experimented with a variation where Transformer was applied at the prediction head level and CBAM was used to refine the input features at the head network, leading to a 4.75% mAP improvement.

The major distinction between previous explorations and this study lies in the proposed architectural method. Previous works primarily used the baseline YOLOv5 models for safety helmet detection [3], without implementing structural improvements to mitigate challenges related to multi-scale, dense, and small object detection. This study addresses those limitations by incorporating CBAM, Transformer, and BiFPN modules into YOLOv5s framework. By conducting the process, this study aims to produce a lightweight, efficient, and highly accurate model capable of overcoming the shortcomings observed in previous implementations.

III. METHODS

This study proposed a modified YOLOv5s [14] algorithm using CBAM [23], Transformer [24], and BiFPN [21] to improve model accuracy as well as multi-scale detection capability in detecting safety helmets. The steps used during the process of the study flow were shown in Fig. 1. First, preprocessing was conducted to divide the image data, which included training, validation, and testing, in addition to converting the label format to YOLO format. Second, modifications are made to the backbone and head networks with CBAM, Transformer, and BiFPN. Third, modeling was performed during the process using the improved architecture

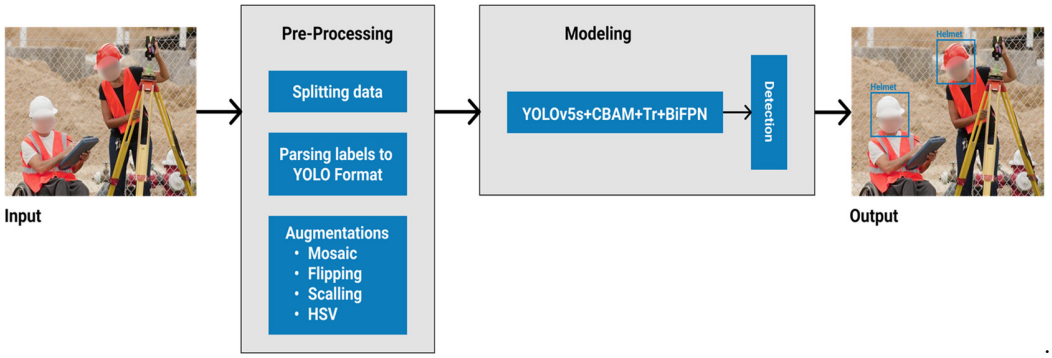

Fig. 1 Proposed methodology's workflow

A. Dataset

The dataset used in this study was a collection of images of construction and manufacturing workers called the Hardhat Dataset, as the samples of the dataset were shown in Fig. 2. During the process, the Hardhat Dataset [25] was obtained from the Harvard Dataverse site shared by Northeastern University China. This dataset, consisting of 7.063 images with a total of 26.633 object annotations, was divided into two classes, namely helmet and head. The annotation format in the dataset was Extensible Markup Language (XML), with the distribution of helmet class totaling 19.852 annotations and head containing 6.781 annotations.

Fig. 2 Sample images of dataset [25]

B. Prepocessing

1) Splitting Dataset

The division of the dataset was performed to adjust the location of the image before it was used in the modeling process, as image data was divided into the train, validation, and test. The division comprised 5,297 training data (75%), 1,059 validation data (15%), and 707 test data (10%). Moreover, each set was placed in a different folder that represented the set.

2) Parsing Dataset

Labels from the dataset originally annotated in XML format were converted to YOLO format to be compatible with YOLO architecture. During the analysis, the center of the object was obtained by calculating the x and y coordinates. The dimension of the bounding box was obtained by calculating the width and height dimensions. Following the process, the center coordinate and bounding box dimensions were normalized by dividing the variables relative to the width as well as height values of the image.

The parsing stage was performed using ElementTree module to retrieve information from XML file. The information retrieved was the image size in length and width, then the bounding box in the form of xmin, ymin, xmax, as well as ymax coordinates, and the last was the class name. Relating to the process, the information was stored in a dictionary and collected into a list dataset. The coordinate value and image size stored in the list dataset were used in the calculation process to get annotations in YOLO format.

3) Augmentation

The dataset used in the modeling process with YOLOv5 was augmented with four methods, namely mosaic, flipping, scaling, and HSV color change. Augmentation in YOLOv5 was applied with mosaic as the main augmentation method used. Mosaic combined four images, namely one original and three additional images randomly selected from the training dataset, which were then randomly rotated and scaled until the figures were finally put together in a grid-like shape. This method was particularly beneficial for helmet detection because the model generated training images with higher object density and variety, effectively improving the ability of the model to recognize small helmet instances in diverse, crowded contexts. The procedure also enriched the training set and improved generalization, making the model more robust in complex environments. The result of the augmentation process during this study was shown in Fig. 3. Based on the mosaic image, there was a probability to apply additional augmentation with the flipping method and color adjustment with HSV.

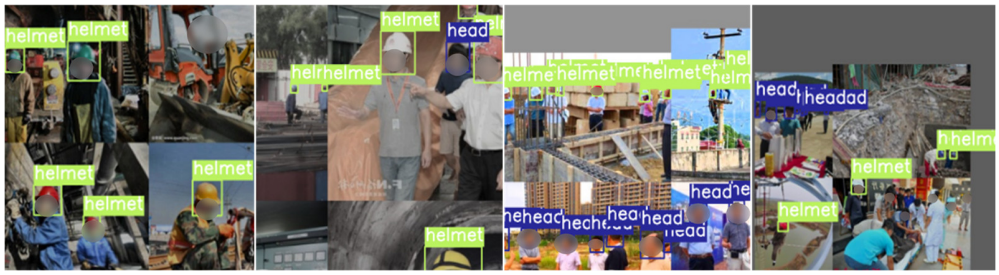

Fig. 3 Sample augmented images

C. Proposed Model

CBAM was an attention module designed to improve the ability of the model to extract important target features on two main dimensions, namely the channel and spatial axes. The attention mechanism efficiently focused on important or suppressed unimportant information. CBAM consisted of two submodules, namely Channel Attention Module (CAM) and Spatial Attention Module (SAM). CAM module focused on the different weights of the channels and multiplied the channels with appropriate weights to prioritize the information on important channels. Moreover, the spatial information from the feature map $F \in R^{C \times H \times W}$ was combined using average-pooling and max-pooling operations to produce two $C \times 1 \times 1$ channel mappings, each of the resulting values was applied in a multi-layer perceptron operation, and the resulting output was summed element-wise. In SAM module, the channel information from the feature map was combined using average-pooling and max-pooling operations to produce two $H' \times W' \times 1$ channel mappings. The final result was obtained by combining two feature maps followed by a 7×7 convolution operation [23].

Transformer [26] architecture was designed based on attentional mechanisms, and the system performed positional encoding of the extracted feature network, which was recombined with the main input vector to produce more refined features. The resulting features were computed and outputted in multi-scale parallel by the decoding process [27]. Each encoder of transformer had two sub-layers, where the first was a multi-head attention layer and the second was fully-connected. Following the discussion, residual connections were applied between sub-layers to help the model learn complex functions more effectively and to mitigate the vanishing gradient problem, enabling deeper architectures to train successfully [22].

BiFPN was a bidirectional architecture that combined top-down and bottom-up pathways. It was designed to enable efficient cross-scale connections and weighted feature fusion for improved multi-scale feature representation [21]. The model combined multi-scale feature information from the backbone network by applying up-sampling and down-sampling operations. It made the feature map resolutions combined and effectively fused information across different scales [28].

This study introduced CTBi-YOLOv5s, an improved version of YOLOv5s designed for safety helmet detection with improvements in both the backbone and head of the network. The backbone was strengthened by incorporating Transformer module, and specifically, C3TR block positioned after SPP (Spatial Pyramid Pooling) layer. This addition leveraged the self-attention mechanism to improve global feature extraction. It improved the ability of the model to understand complex visual patterns.

Further improvements were made in the head architecture, where CBAM, Transformer layers, and BiFPN were introduced. CBAM was incorporated before the feature map combining stage to show critical spatial and channel-wise features. It improved the ability of the network to distinguish object shapes and boundaries. Additionally, Transformer blocks replaced the standard C3 modules in the prediction head to strengthen feature representation and contextual understanding. These incorporated improvements aimed to boost the accuracy and robustness of the model, specifically in challenging construction site environments.

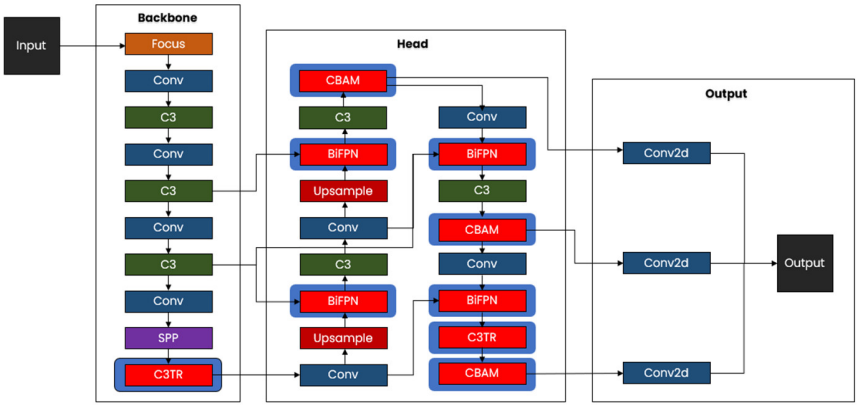

Fig. 4 Proposed CTBi-YOLOv5s

The use of Transformer in this section was intended to use the self-attention mechanism that allowed the model to analyze the relationship of various feature maps extracted from the backbone network. After the process, BiFPN was applied to the head architecture network at the medium resolution level. In this network structure, the feature map

from the previous layer was combined with two other feature maps. The union included one from the backbone having a mid-resolution feature map (P4), and two from layer 6, which was previously where this layer was a higher-level feature (P3) [20]. During the process, the bottom-up and top-down mechanisms of BiFPN allowed the model to effectively combine features from different scales for improving the feature representation of the object. The application of BiFPN had the potential to overcome the multi-scale problem of YOLOv5. The diagram of the total framework of CTBi-YOLOv5s model was shown in Fig. 4.

D. Evaluation Matrix

Evaluation was used to measure the performance of the model during the process. Model evaluation was performed using several matrices, namely mAP, recall, precision, and F1 score. mAP matrix was the average value of the model in predicting objects in various categories. Additionally, it was used to assess the effectiveness of the model in object detection, and the following was the formula for the matrix (1).

$$mAP = \sum_{q=1}^{Q} \frac{AveP\,(q)}{Q} \quad (1)$$

Precision was a matrix that measured the accuracy of the model in making positive predictions. Meanwhile, recall measured the accuracy of the model in making predictions that were positive. The F1 score was used to evaluate the model based on precision and recall values. The following were the formulas of precision (2), recall (3), and F1 score (4).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 } Score = 2\ x\ \frac{Precision\ x\ Recall}{Precission+Recall} \quad (4)$$

IV. RESULTS

Experiments were conducted using the Hardhat dataset with the basic and Improved YOLOv5s algorithm. The device used in the modeling process was Google Collaboratory by using the T4 GPU. Additionally, parameters from [3] were used, consisting of batch size 20, epoch 50, and image size 416 x 416. The model was evaluated using precision, recall, F1 score, and mAP. During the process, systems that had been modified with BiFPN, CBAM, and Transformer were compared with YOLOv5s base model. Table 1 showed the comparison of evaluation results after modeling.

TABLE 1
RESULT OF THE EXPERIMENT

Model	Precision (%)	Recall (%)	F1 Score	mAP@0.5
YOLOv5s	93.9	87.1	0.90	0.936
YOLOv5s + BiFPN	94	94.2	0.94	0.972
YOLOv5s + CBAM	94.8	92.7	0.94	0.97
YOLOv5s + Tr	95.1	93.4	0.94	0.972
YOLOv5s + CBAM + BiFPN	94.2	93.3	0.94	0.97
YOLOv5s + Tr + BiFPN	95	92.9	0.94	0.973
YOLOv5s + Tr + CBAM	94.8	9.22	0.94	0.971
YOLOv5s + Tr + CBAM + BiFPN	94.1	93.3	0.94	0.973

*Tr represented Transformer module

A significant contribution of this study was the determined improvement achieved by modifying YOLOv5s architecture with BiFPN, CBAM, and Transformer components. Modeling results showed that these modifications provided a substantial improvement in detection performance compared to the base model. The original YOLOv5s achieved 93.9% precision, 87.1% recall, 0.90 F1 score, and 93.6% mAP@0.5. Through the improvement of this study, the F1 score improved by up to 4%, while precision increased by 0.1%, 0.7%, and 1.2%. Recall improved by 7.1%, 5.6%, and 6.3%, and mAP@0.5 rose by 3.6%, 3.4%, as well as 3.6%, depending on the combination of modules applied.

The combination of CBAM and BiFPN led to 94.2% precision, 93.3% recall, and a mAP@0.5 of 97%, signifying a strong synergistic effect between the two improvements. The joining of Transformer and BiFPN achieved 95% precision, 92.9% recall, and 97.3% highest mAP@0.5, marking the best precision among the tested models with a slightly less balanced recall. During the analysis, the combination of Transformer and CBAM produced 94.8%

precision, 92.2% recall, and mAP@0.5 of 97.1%. The incorporation of all three modules (Transformer, CBAM, and BiFPN) produced the best total performance, achieving 94.1% precision, 93.3% recall, and 97.3% mAP@0.5.

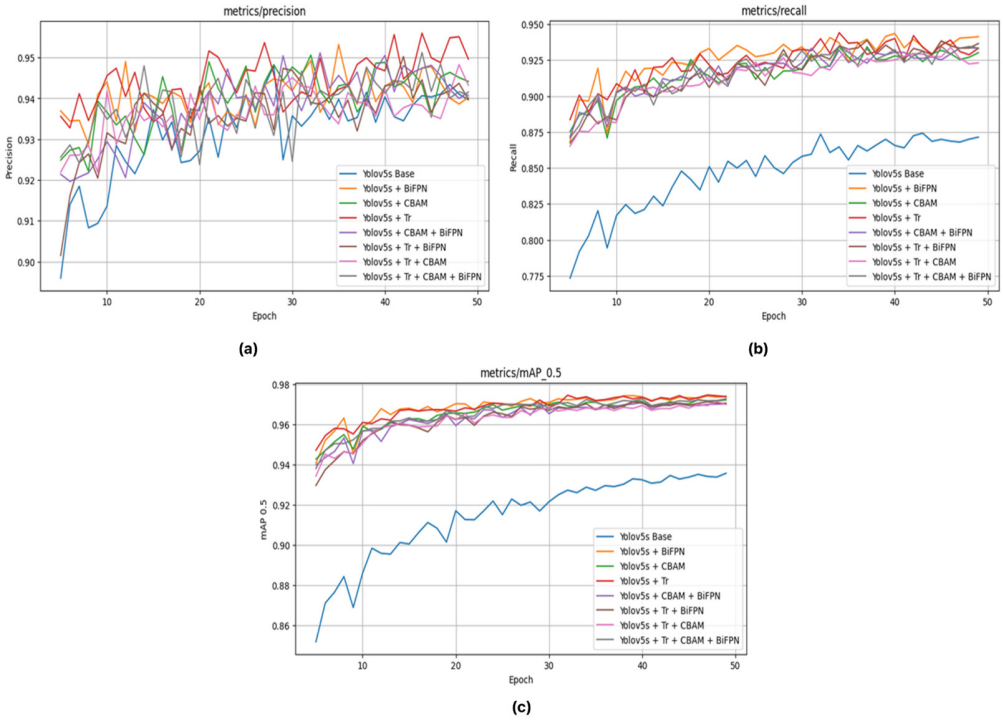

Fig. 5 Model progress during training (a) Precision; (b) Recall; (c) mAP@0.5

The training progress of the model during this study was shown in Fig. 5. The improvements, particularly in recall and mAP, were significant compared to the base YOLOv5s model, showing the effectiveness of the proposed architectural modifications. The addition of Transformer and CBAM modules influenced the complexity of the model. The basic YOLOv5s model had 166 layers, which increased to 176 through the addition of a Transformer and further to 202 with the incorporation of CBAM. However, these modifications led to significant improvements in precision, recall, F1 score, and mAP.

TABLE 2
COMPARISON OF SIZE, TRAINING TIME, AND MODEL INFERENCE SPEED

Model	Size	Time	FPS
YOLOv5s	14.4MB	55min 57s	104
YOLOv5s + BiFPN	14.5MB	48min 09s	105
YOLOv5s + CBAM	14.5MB	53min 17s	70
YOLOv5s + Tr	14.4MB	48min 38s	71
YOLOv5s + CBAM + BiFPN	14.6MB	46min 55s	79
YOLOv5s + Tr + BiFPN	14.5MB	54min 47s	54
YOLOv5s + Tr + CBAM	14.5MB	50min 48s	55
YOLOv5s + Tr + CBAM + BiFPN	14.6MB	48min 02s	58

Modifications to the architecture affected the training process and the speed of model inference. Table 2 showed the model inference speed in Frames per Second (FPS), size, and training time. FPS measured how many frames or images the model processed per second during inference. Based on the ability, the modified models led to larger model sizes since the replicas required slightly faster training time. The improvements in detection accuracy came with a drawback, where the inference speed dropped due to increased computational complexity. However, this drawback was justified by the substantial advances in detection performance for complex construction site environments. Compared to the previous study conducted by [3], the improvements applied provided better results in terms of accuracy and detection. Table 3 showed the comparison between the previous study and the proposed model.

TABLE 3
COMPARISON OF THE PREVIOUS STUDY AND THE PROPOSED MODEL

Model	Precision (%)	Recall (%)	F1 Score	mAP@0.5
YOLOv5n	93.4	87.6	0.904	0.942
YOLOv5s	93.6	89.9	0.917	0.953
YOLOv5m	94.8	90.4	0.925	0.955
YOLOv5l	94.9	94.9	0.925	0.956
YOLOv5x	93.9	93.9	0.925	0.958
CTBi-YOLOv5s (Proposed)	94.1	93.3	0.94	0.973

V. DISCUSSION

Multi-scale detection was a challenging task in safety helmet detection due to significant variation in object size, which arose from the varying distances between workers and the camera in real-world environments. In this study, most safety helmets and human heads appeared relatively small in the image frame, making detection particularly difficult. The size of helmet was not fixed but relative to the general image resolution, and in a single frame, objects of interest varied considerably in scale. These conditions showed the need for robust multi-scale feature learning. A previous study by [6], which introduced the HardHat dataset and also focused on helmet detection, was conducted under experimental conditions different substantially from those in this analysis.

A trend surfaced when this study compared each improvement against YOLOv5s baseline based on the results by [3]. As the original model already delivered high precision (93.9%) and respectable mAP@0.5 (93.6%), its recall (87.1%) lagged, showing that up to 13% of objects were missed. Introducing BiFPN dramatically narrowed this gap, improving recall to 94.2% and raising mAP@0.5 by 3.6% with only a marginal 0.1% gain in precision. This shows that the weighted feature fusion of BiFPN allows the detector to capture a broader variety of object scales without sacrificing accuracy.

CBAM module, which applied attention sequentially across channels and spatial locations, shifted the balance slightly toward precision, as precision rose to 94.8% and recall to 92.7%, improving the focus of the model on truly relevant regions. mAP@0.5 increase of 3.4% signified that feature recalibration sharpened the judgment of the network with a tad less sensitivity than BiFPN. A Transformer head ("Tr") pushed precision even higher to 95.1%, while raising recall to 93.4% and mAP@0.5 by 3.6%. This followed the hypothesis that self-attention layers helped the network to incorporate long-range dependencies and contextual relationships, making for fewer false positives as well as false negatives.

The combination of modules in this study showed complementary effects during the process. Joining CBAM with BiFPN produced a model balanced at 94.2% precision, 93.3% recall, and 97.0% mAP@0.5, signifying that feature fusion as well as attention synergized to both broaden detection coverage as well as refine focus. During the process, pairing Transformer and BiFPN achieved the highest mAP@0.5 of 97.3% while maintaining 95.0% precision as well as 92.9% recall. Multi-scale ability of BiFPN incorporated seamlessly with representational depth of self-attention, maximizing mean average precision. The combination of Transformer and CBAM also produced strong achievements (94.8% precision, 92.2% recall, 97.1% mAP@0.5), but the slightly reduced recall showed that overlapping attention mechanisms might have led to diminishing returns in sensitivity.

Incorporating all three modules produced an exceptionally well-rounded detector, including 94.1% precision, 93.3% recall, and the shared top mAP@0.5 of 97.3%. This full-stack model reached near-optimal mAP while preserving strong precision and recall, confirming that multi-scale fusion as well as dual attention pathways were layered without conflict. In practical terms, when maximum mAP was the priority and computational overhead was acceptable, the triple-module YOLOv5s was outstanding. For scenarios where inference speed or resource constraints matter more, a single–module improvement, specifically BiFPN, delivered most of the recall and mAP achievements at lower complexity.

Table 2 showed that while all improved YOLOv5s variants slightly increased model size and added computational overhead, Transformer + CBAM + BiFPN combination delivered the most favorable drawback. In this study, the baseline YOLOv5s (14.4 MB) trained in 55 min 57 s and achieved 104 FPS. Adding BiFPN alone (14.5 MB) actually reduced training time to 48 min 09 s and pushed FPS to 105, a negligible size increase for faster learning as well as inference. Consequently, CBAM or Transformer personally cut training time by 2–7 minutes and incurred a steep drop in FPS to 70 and 71, signaling that attention mechanisms added a measurable runtime cost. When all three modules were combined, the model increased only to 14.6 MB, yet training completed in just 48 min 02 s, faster than the baseline, and inference still ran at a robust 58 FPS, higher than the 30 FPS real-time threshold for video applications. Tr + CBAM + BiFPN variant appeared as the best selection, supported by its top mAP@0.5 performance. It maximized detection accuracy while maintaining real-time inference speed and only a modest increase in footprint.

Table 3 showed that the modified model significantly outperformed previous studies, achieving an impressive mAP of 97.3%, compared to 95.3% for the previous YOLOv5s model and 95.8% for the highest mAP for YOLOv5x model. This improvement was further shown in Figure 6, where the inference results of the base and improved models were compared. The blue circles signified missed detections by the base YOLOv5s model, which the modified model successfully detected, showing its ability to accurately identify objects in complex scenes. These improvements reduced false detections and signified the effectiveness of the modifications in improving multi-scale object recognition, particularly in challenging environments such as construction sites with varying object sizes as well as occlusions. This study incorporated three major components, namely CBAM, Transformer, and BiFPN, to overcome the limitations of the baseline YOLOv5s model in detecting helmets in cluttered as well as dynamic construction environments.

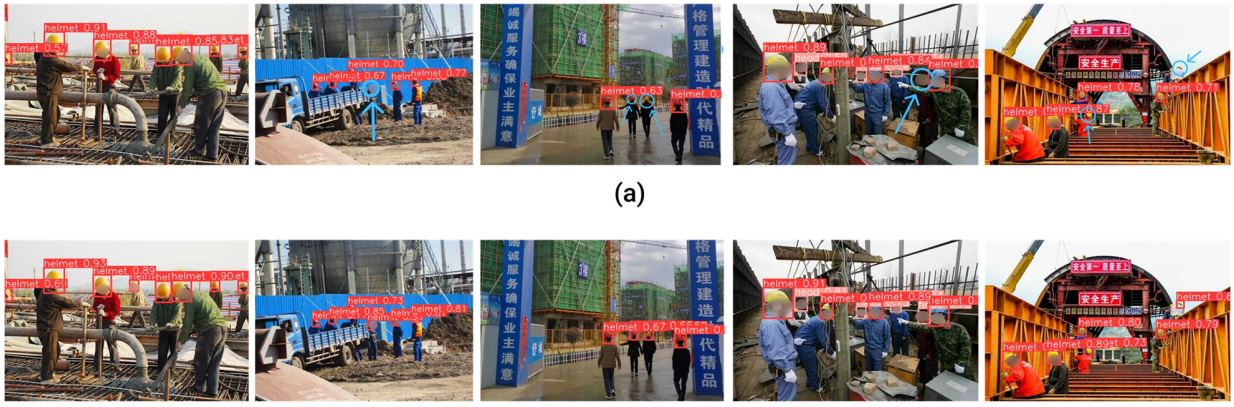

Fig. 6 Comparison of helmet detection performance before and after model improvement. (a) Detection results using the baseline YOLOv5s model. (b) Detection results using the proposed CTBi-YOLOv5s model.

CBAM improved the ability of the model to focus on the most relevant spatial and channel-wise features, improving detection accuracy in visually complex scenes. Transformer module introduced a global self-attention mechanism, allowing the model to capture long-range dependencies and better interpret contextual relationships, particularly useful in crowded settings. Meanwhile, BiFPN supported efficient multi-scale feature fusion, which was essential for accurately detecting helmets of various sizes and at different distances. These components form a lightweight yet powerful architecture that significantly improved detection performance, maintaining the speed and efficiency required for real-time deployment in construction site safety monitoring.

VI. CONCLUSIONS

In conclusion, this study introduced a novel YOLOv5s variant that incorporated three complementary architectural components, namely CBAM, Transformer-based self-attention, and BiFPN, to simultaneously improve spatial and channel-wise attention, capture long-range dependencies, as well as perform efficient multi-scale feature fusion. This study conducted a comprehensive evaluation across eight model configurations, assessing detection performance (precision, recall, F1 score, and mAP@0.5), model size, training time, as well as inference speed. By quantifying these metrics, the analysis provided practical deployment understanding and drawback analyses modified for real-time helmet detection in construction site environments.

The results showed that the combined Tr + CBAM + BiFPN model achieved the best total performance, improving mAP@0.5 from 93.6% to 97.3%. Precision was also improved to 94.1%, recall to 93.3%, and F1 score to 0.94%, while maintaining a compact footprint (14.6 MB), reducing training time by up to 14% (48 min 02 s vs. 55 min 57 s), as well as delivering real-time inference at 58 FPS. BiFPN produced the greatest recall improvement to 94.2% and Transformer head delivered the highest precision of 95.1%, but only the full incorporation balanced peak accuracy with practical speed as well as memory requirements, making it the recommended selection for deployment.

The analysis should recommend for future studies a probabilistic exploration of module combinations through neural architecture search or Bayesian optimization to identify even more lightweight yet effective variants, potentially in appearing frameworks such as YOLOv8-nano, YOLOv7-tiny, or EfficientDet-D0. Methods such as knowledge distillation, structured pruning, and post-training quantization should be applied to compress the model further without sacrificing accuracy, while conditional computation or dynamic gating could adaptively activate attention as well as

fusion modules based on scene complexity. Deploying and benchmarking these optimized architectures across diverse edge and embedded platforms would be crucial to ensure robust, low-latency performance in real-world, resource-constrained environments.

Author Contributions: *Tio Dharmawan*: Reviewed the research. *Danang Setiawan*: Conducted the research, Wrote and revised the manuscript. *Muhammad Arief Hidayat*: Conducted a review related to manuscript writing. *Vandha Pradwiyasma Widartha*: Conducted a review related to manuscript writing.

All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from the Data Science Research Group, Computer Science Faculty, University of Jember.

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

Data Availability: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/7CBGOS

Informed Consent: Informed Consent was obtained, and a detailed explanation was presented in the Methods section.

Institutional Review Board Statement: Not applicable.

Animal Subjects: There were no animal subjects.

ORCID:
Tio Dharmawan: https://orcid.org/0000-0003-3692-2691
Danang Setiawan: https://orcid.org/0009-0004-1070-7381
Muhammad Arief Hidayat: https://orcid.org/0009-0000-4195-188X
Vandha Pradwiyasma Widartha: https://orcid.org/0000-0001-9790-7597

REFERENCES

[1] R. L. Mathis and J. H. Jackson, Human Resource Management: Manajemen Sumber Daya Manusia. 2002.

[2] L. M. Azzahri and K. I. Ikhwan, "Hubungan Pengetahuan Tentang Penggunaan Alat Pelindung Diri (APD) Dengan Kepatuhan Penggunaan Apd Pada Perawat Di Puskesmas Kuok," *Prepotif: Jurnal Kesehatan Masyarakat*, vol. 3, no. 1, pp. 50–57, 2023, doi: 10.31004/prepotif.v3i1.442.

[3] M. U. Kisaezehra, M. A. Farooq, A. Bhutto, and A. K. Kazi, "Real-time safety helmet detection using yolov5 at construction sites," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 911–927, 2023.

[4] N. Kwak and D. Kim, "Detection of Worker's Safety Helmet and Mask and Identification of Worker Using Deeplearning," *Computers, Materials & Continua*, vol. 75, no. 1, pp. 1671–1686, 2023, doi: 10.32604/cmc.2023.035762.

[5] Y. Said *et al.*, "AI-Based Helmet Violation Detection for Traffic Management System," *Computer Modeling in Engineering & Sciences*, vol. 141, no. 1, pp. 733–749, 2024, doi: 10.32604/cmes.2024.052369.

[6] L. Wang, L. Xie, P. Yang, Q. Deng, S. Du, and L. Xu, "Hardhat-Wearing Detection Based on a Lightweight Convolutional Neural Network with Multi-Scale Features and a Top-Down Module," *Sensors*, vol. 20, no. 7, p. 1868, Mar. 2020, doi: 10.3390/s20071868.

[7] W. Chen, M. Liu, X. Zhou, J. Pan, and H. Tan, "Safety Helmet Wearing Detection in Aerial Images Using Improved YOLOv4," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 3159–3174, 2022, doi: 10.32604/cmc.2022.026664.

[8] H. Li, D. Wu, W. Zhang, and C. Xiao, "YOLO-PL: Helmet wearing detection algorithm based on improved YOLOv4," *Digit Signal Process*, vol. 144, p. 104283, Jan. 2024, doi: 10.1016/j.dsp.2023.104283.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[10] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016, [Online]. Available: https://arxiv.org/abs/1612.08242v1

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018, [Online]. Available: https://arxiv.org/abs/1804.02767

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal Speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020, [Online]. Available: https://arxiv.org/abs/2004.10934

[13] A. R. Muhammad, H. P. Utomo, P. Hidayatullah, and N. Syakrani, "Early Stopping Effectiveness for YOLOv4," *Journal of Information Systems Engineering and Business Intelligence*, vol. 8, no. 1, pp. 11–20, 2022, doi: 10.20473/jisebi.8.1.11-20.

[14] G. Jocher, K. Nishimura, T. Mineeva, and R. J. A. M. Vilariño, "yolov5," 2020.

[15] Shan Du, M. Shehata, and W. Badawy, "Hard hat detection in video sequences based on face features, motion and color information," in *2011 3rd International Conference on Computer Research and Development*, IEEE, Mar. 2011, pp. 25–29. doi: 10.1109/ICCRD.2011.5763846.

[16] K. Shrestha, P. P. Shrestha, D. Bajracharya, and E. A. Yfantis, "Hard-Hat Detection for Construction Safety Visualization," *Journal of Construction Engineering*, vol. 2015, pp. 1–8, Feb. 2015, doi: 10.1155/2015/721380.

[17] T. Mahendrakar and others, "Performance Study of YOLOv5 and Faster R-CNN for Autonomous Navigation around Non-Cooperative Targets," in *2022 IEEE Aerospace Conference (AERO)*, Big Sky, MT, USA, 2022, pp. 1–12. doi: 10.1109/AERO53065.2022.9843537.

[18] H. Chen, Z. Chen, and H. Yu, "Enhanced Yolov5: An efficient road object detection method," *Sensors*, vol. 23, no. 20, p. 8355, 2023, doi: 10.3390/s23208355.

[19] F. Zhou, H. Zhao, and Z. Nie, "Safety Helmet Detection Based on YOLOv5," in *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, Shenyang, China, 2021, pp. 6–11. doi: 10.1109/ICPECA51329.2021.9362711.

[20] H. Li, L. Shi, S. Fang, and F. Yin, "Real-Time Detection of Apple Leaf Diseases in Natural Scenes Based on YOLOv5," *Agriculture*, vol. 13, no. 4, p. 878, 2023, doi: 10.3390/agriculture13040878.

[21] M. Tan, R. Pang, and Q. V Le, "EfficientDet: Scalable and efficient object detection," *arXiv preprint arXiv:1911.09070*, 2020, [Online]. Available: https://arxiv.org/abs/1911.09070

[22] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, 2021, pp. 2778–2788. doi: 10.1109/ICCVW54120.2021.00312.

[23] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Computer Vision – ECCV 2018*, vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11211. , Springer, Cham, 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[24] A. Dosovitskiy *et al*., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[25] L. Xie, "Hardhat," 2019. [Online]. Available: https://doi.org/10.7910/DVN/7CBGOS

[26] A. Vaswani and others, "Attention is All you Need," *Neural Information Processing Systems*, 2017, doi: 10.48550/arXiv.1706.03762.

[27] J. Yao, X. Fan, B. Li, and W. Qin, "Adverse Weather Target Detection Algorithm Based on Adaptive Color Levels and Improved YOLOv5," *Sensors*, vol. 22, no. 21, p. 8577, 2022, doi: 10.3390/s22218577.

[28] Y. Guo, P. Regmi, Y. Ding, R. B. Bist, and L. Chai, "Automatic detection of Brown Hens in cage-free houses with Deep Learning Methods," *Poult Sci*, vol. 102, no. 8, p. 102784, 2023, doi: 10.1016/j.psj.2023.102784.