IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 April 2024, accepted 28 April 2024, date of publication 2 May 2024, date of current version 10 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396224

RESEARCH ARTICLE

Improved YOLOv8 Model for a Comprehensive Approach to Object Detection and Distance Estimation

ZU JUN KHOW, YI-FEI TAN, HEZERUL ABDUL KARIM, (Senior Member, IEEE), AND HAIRUL AZHAR ABDUL RASHID, (Senior Member, IEEE)

Faculty of Engineering, Multimedia University, Cyberjaya 63100, Malaysia

Corresponding author: Yi-Fei Tan (yftan@mmu.edu.my)

ABSTRACT The rapid advancements in deep learning have revolutionized the field of computer vision. However, despite the significant progress in computer vision, there remains a scarcity of research focused on utilizing this technology for distance estimation. Exploring such studies can bring immense convenience to people, especially in applications like anomaly object detection. On that account, this research proposes an improved detection model based on You Only Look Once version 8 (YOLOv8) namely YOLOv8-CAW, which is capable of both detecting target objects and accurately calculating their distances. The proposed method involves incorporating the Coordinate Attention and Wise-IoU into the YOLOv8 network, enhancing the detection accuracy. Combined with the distance estimation algorithm, results in a comprehensive output that includes both detection results and calculated distances. At the end of the experiment, a substantial improvement in performance metrics was observed, the model achieved increases in recall (0.4%), precision (2.2%), and Mean Average Precision (mAP) (1.5%) within the 0.5 to 0.95 threshold range, while maintaining inference speeds similar to the baseline model in PASCAL VOC dataset. Besides that, distance estimation achieved an approximate average accuracy of 90% which shows the results are highly encouraging and promising. The successful integration of computer vision and distance estimation opens new possibilities for practical applications, showcasing the potential of this approach in real-world scenarios.

INDEX TERMS You only look once (YOLO), deep learning, object detection, distance estimation, attention module, loss function.

I. INTRODUCTION

Computer vision has undergone a revolutionary transformation with the rapid advancements in deep learning. The application of deep learning techniques has significantly enhanced the capabilities of computer vision systems, enabling them to perform tasks with unprecedented accuracy and efficiency. With creative and reasonable use, computer vision techniques can even complete tasks that humans cannot. To date, computer vision has been developed with many interesting applications, authors in [1] has completed research that uses computer vision techniques to calculate an object's width and height, [2] has integrated the processing of images from visual and infrared cameras for forest fire detection, [3] has used computer vision techniques for violence detection in video, and yet, all of these are just a tiny fraction of the potential that computer vision has.

Given the potential power of computer vision techniques, this research proposes a distance estimation technique that adopts YOLOv8, developed by Jocher et al. [4] as its foundation. In real life, people need to estimate distances in various situations, For both industrial and daily needs. For example, in a pipeline cleaning company, if the dirt inside can be discovered knowing the exact location, it could help speed up the cleaning process and thus maximize productivity. Besides that, if the distance between the cars can be detected, and an early signal could be sent to the drivers, it may help reduce the risk of exposure to accidents. These examples have

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

inspired us to carry out research to better estimate the distance between objects by using the computer vision principle.

In this paper, we report a modified improved YOLOv8 architecture namely YOLOv8-CAW for object detection that achieves better results on a variety of devices is proposed, by modifying the original YOLOv8 architecture. Firstly, the YOLOv8-CAW model is integrated with the Coordinate Attention (CA) module proposed by Hou et al. [5] which aimed to improve the model accuracy without significantly increasing the model size. Secondly, the original YOLOv8 Complete Intersection over Union (C-IoU) loss function developed by Zheng et al. [6] is replaced by the Wise-IoU (WIoU) loss function developed by Tong et al. [7] which led to a better training convergence. For distance estimation detection, the ratio calculation method is used to compare the average size of the object in the real world with the size of the object when it is captured in computer vision to estimate the distance between the object and the camera, details will be explained in Section III. The rest of the paper is organized as follows, Section II reviews the state-of-the-art research on object detection and distance estimation. Section III describes the proposed method in detail, while Section IV presents the experiment setup, Section V will discuss the experiment result, lastly, Section VI will give a short summary of our findings.

II. RELATED WORK

The discussion is structured into two distinct sections: (A) Concentrating on object detection techniques and their real-world applications. (B) Expanding on the implementation of the distance estimation component.

A. OBJECT DETECTION

In the current object detection research direction, the YOLO object detection model or other object detection models, as an emerging topic, has been subject to repeated experimentation and enhancement by researchers. Each researcher has their own set of methods for improving the overall performance and capabilities of the model. Researchers in [8] proposed that, rather than solely focusing on the depth of the model, enhancing the model's performance can be better achieved by enabling it to understand the contextual relationship between shallow and deep layers. In the relative experiment, researchers combined YOLOv4 with PANet [9] where the shallow network passes the information to the deeper network for the feature fusion and Squeeze and Excitation (SE) block [10] that enhance model performance by automatically learning and emphasizing important features while reducing redundancy and increasing network expressiveness. This experiment yielded a Mean Average Precision (mAP) of 89.63%, surpassing the original YOLOv4 model by 2.86%. In the same study of YOLOv4, authors in [11] leveraged an adaptive context module to balance foreground and background features to obtain global contextual information. They also introduced a balanced prediction layer method to mitigate feature-level imbalances and an anti-congestion network for finer-grained detection. Additionally, authors utilized a tailored heterogeneous cross-entropy loss during training to improve target discrimination across categories. Separate study of the YOLOv4 model. Researchers in [12] successfully addressed the multi-scale detection issue by employing Spatial Pyramid Pooling (SPP) [13] in conjunction with YOLOv4. SPP effectively divides the input feature map into multiple grids of varying sizes and applies pooling operations to each grid. This approach enables the model to extract features at different scales, enhancing its ability to identify smaller objects with greater accuracy. The resulting performance is highly satisfactory. In addition to SPP, researchers explored alternative methods to tackle the small-scale detection challenge. For instance, in the research [14] the authors employed YOLOv8 and a combination of 1 × 1 convolutions for dimensionality reduction and 3 × 3 convolutions for down sampling. This strategy effectively preserves contextual information during the feature extraction process, facilitating a more comprehensive fusion of features between shallow and deeper layers. Additionally, a technique that allows each layer to receive crucial information from all preceding layers is implemented, enabling the network to capture a more complete representation of the contextual information. Experimental results demonstrated that this approach outperforms the baseline model on all datasets.

Research [15] leveraged YOLOv8 as the basis for the experiments and integrated various techniques to tackle the challenge of detecting small-scale objects. Authors utilized GSConv [16] to reduce computational complexity and implemented Content-Aware Reassembly of Features (CARAFE) [17] to extract and utilize contextual information more efficiently. Additionally, they enhanced the model's performance by replacing the original Spatial Pyramid Pooling Fusion (SPPF) module with their implementation and introduced an object detection layer to enhance the fusion of shallow and deep feature maps. These modifications led to a model that demonstrated superior performance in small-scale object detection while maintaining competitive computational efficiency. On the other hand, authors in [18] addressed the issue by incorporating their network with a channel attention (CA) module and designing a new backbone for YOLOX, integrating CSPDarknet with the inverted residual block. This combination enables the model to be more sensitive to small-scale objects while preserving lower computational complexity. And unexpectedly, YOLOX was mentioned, which is also noteworthy. Researchers in [19] tried different methods to enhance YOLOX performance, the researchers introduced a novel neck design and a double residual branch structure in the detection head, to enhance the detection accuracy and efficiency. Back to the small-scale detection issue, authors in [20] also tackled the small-scale object detection challenge by integrating YOLOv8 with the WIoU loss function and the BIFORMER [21] attention mechanism. This approach achieves more flexible computation resources allocation and content-aware attention by initially filtering out irrelevant key-value pairs at the coarse-grained regional level. Additionally, the authors added two new detection heads to YOLOv8. Researchers in [22] explored a different method to detect the small objects. They introduced the Context Attention Block for multi-scale feature localization. Additionally, it enhances feature extraction and accelerates detection performance without increasing model complexity by modifying the C2f block. Spatial Attention was also modified to augment model performance. The experiment resulted in 0.9% higher than the original model in mAP compared to original model. Meanwhile, researchers in [23] addressed small object issues without using the YOLO models. They experimented with SSD (Single Shot MultiBox Detector) [24] by improving the backbone network with multi-scale feature fusion and an attention mechanism. Subsequently, they boosted shallow network feature extraction for better small object recognition. Afterward, they utilized RFB (Receptive Field block) [25] to widen object receptive fields and gather richer semantic information. Lastly, an attention mechanism was incorporated to emphasize important object features and suppress irrelevant information.

Similar to [23], the study in [26] also focused on semantic information retrieval but with YOLOv5 model. Authors addressed semantic information issue by introduced a novel object detection network based on a large kernel convolutional neck network to enhance semantic feature capture. Additionally, a Vast Receptive Field Attention mechanism was constructed to increase receptive field. Meanwhile, authors in [27] addressed the semantic gap issue using YOLOv5 by merging two adjacent low-level features from the Feature Pyramid Network (FPN) and gradually integrating higher-level features to avoid a larger semantic gap. Besides that, authors in [28] introduced a module that reinjects high-resolution details from shallow features and prioritizes informative channels. Another module expands the context for each pixel and incorporates global image information to avoid semantic information loss.

The researchers in [29] proposed a method that can reduce YOLOv5 model parameters and make the model leverage global contextual information to increase model performance. The authors used the Transformer block [30] to empower the model to effectively utilize global contextual information. Additionally, they integrated rep modules [31] and Stem modules [32] to reduce computational complexity. These techniques yielded promising results, demonstrating significant improvements in model performance. Research in [33] conducted a similar study; however, their approach differed in using the Vision Transformer [34] to capture global contextual information and integrating the Convolutional Block Attention Module (CBAM) [35] to enhance feature expression capabilities for foreign object detection. In a separate study reported in [36], a distinct research question on YOLOv8 was pursued. The authors suggested that addressing dataset characteristics involves considering a variety of factors, including non-linear characteristics. Their proposed method aimed to investigate the relationships involving these non-linear characteristics. As part of their approach, the authors replaced the SILU activation function [37], with the MISH activation function [38]. This transition allowed for the use of a more adaptable activation function suitable for handling complex images. Additionally, the researchers integrated a DCF module to aggregate low-level features from the dataset. The outcomes of this study were indeed promising. Authors yielded promise of addressing non-linear dataset characteristics and yielding significant improvements in their research outcomes. Research in [39] identified various issues with the YOLOv5 model. In response to the lack of available snow datasets, authors endeavored to address this gap by creating a novel real-world snowy object detection dataset. Subsequently, they employed an unsupervised learning approach to categorize the snowy dataset into four levels of difficulty. Furthermore, they enhanced the YOLOv5 model by introducing a lightweight modification, incorporating a novel Cross Fusion module.

Authors in [40] focus on reducing the parameters of the YOLOv5 model to enhance computation speed. In their experimental work, the authors employed a decoupled head approach to separate the classification and localization branches within YOLOv5, thereby accelerating the training process. This enhancement became particularly significant as the original C3 model in YOLOv5 necessitates five convolutional operations. The authors introduced their self-proposed C3-faster block, which restructures the original C3 by reducing the number of convolutions from 5 to 3, effectively conserving computational resources. Furthermore, the authors incorporated the WIoU loss function to enhance detection accuracy, resulting in an impressive accuracy rate of 97.1%, compared to the original model's 93.4%. Same goes for reducing YOLOv5 model parameters, authors in [41] incorporated a CA module and adjusted the original YOLOv5's C3 module to reduce model parameters. In contrast, researchers in [42] aimed to reduce the model size with YOLOv8. They replaced the original YOLOv8 Darknet-53 backbone with FasterNet-T0 [43], significantly reducing the model's parameters. Additionally, they added a new detection head to enhance the sensitivity to small objects. Furthermore, researchers incorporated a CA module to improve model performance while minimizing changes to the parameter structure. Following experiments, researchers achieved a 0.5% higher mAP compared to the original model on the PASCAL VOC [44] dataset, while also significantly reducing the parameter size.

In a similar pursuit of a lightweight model, researchers in [45] proposed a different method to reduce the YOLOv8 model parameters while enhancing performance. They merged YOLOv8 with Context GuidedNet [46] to capture contextual features and Res2Net [47] to improve the model's ability to learn deep features while minimizing the impact on model parameters. The researchers also enhanced the model's ability to handle the feature pyramid by restructuring the model and incorporating the WIoU loss function to manage problematic samples.

Furthermore, studies have focused on reducing parameters in models like YOLOv4-Tiny. In [48], authors redesigned the network structure and proposed a Trident-FPN that combines three scales of detection head from YOLOv4-Tiny and incorporates pooling feature augmentation to generate deeper semantic features. This approach effectively obtains semantic information while keeping computational costs low, similar study happened to [49] also. Other than YOLO model augmentation, authors in [50] introduced a dual-path network for real-time object detection. This network uses a lightweight attention mechanism to extract both high-level semantic features and low-level object details efficiently. It also incorporates self-proposed 'Lightweight Self-Correlation (LSCM)' and 'Cross-Correlation (LCCM)' modules to capture global interactions and dependencies among scale features. Authors in [51] also experimented enhance model performance without using YOLO series model. They proposed a module that combines features from low to high level scales, along with an LNblock to improve spatial information extraction ability with lower computational costs. It is noteworthy that not all researchers aim to reduce model parameters by changing the model structure. In [52], researchers introduced the Effective Receptive Field (ERF) module to expand the network's receptive field and optimize path aggregation network structure in detectors for improved accuracy while reducing model parameters for Unmanned aerial vehicle (UAV) deployment.

Researchers in [53] tried to address a unique issue. In recent years, the fisheye camera has become more popular, but the images captured by fisheye camera are often distorted which brings a significant challenge to the field of object detection since it typically works better with regular images. To tackle this issue, they proposed a novel ''Max Pooling and Ghost's Downsampling'' module for extracting the feature from distorted images and an ''Average Pooling and Ghost's Downsampling'' module for acquiring rich global information. Meanwhile, they also modified the original C2f module with SE block to acquire richer gradient flow information about the features. The same technique is also applied to the SPPF module in YOLOv8 to improve the model's ability to detect distorted images. In addition to addressing the issue of distorted images, researchers also experimented model efficiency on the regular image with MS-COCO [54] dataset and resulting in 1.7% higher average precision compared to the original model.

It is worth mentioning that recent work on Neural Architecture Search (NAS) for object detection has also been rapidly evolving. In studies [55], [56], researchers have highlighted the over-reliance of current object detection methodologies on researchers' expertise and knowledge in constructing neural networks. Hence, both researchers had conducted research on NAS with different approaches. Researchers in [55] endeavored to reduce NAS computational cost by proposing a new Early Exit Population Initialisation (EE-PI) algorithm. This algorithm filters out networks and replaces those that surpass a certain threshold with models having fewer parameters. Researchers in [56] argued that recent research on NAS techniques only focused on the backbone or FPN, lacking study on YOLO network which has a more efficient detector head. By referencing the YOLOv5 network, these researchers proposed a framework that can jointly search for the architectures of the backbone and FPN.

Several studies have also explored different strategies to enhance object detection models. For example, authors in [57] addressed complex background issues using ensemble learning, while authors in [58] employed a line encoding method that encodes bounding boxes into the top-left corner and the bottom-right corner, then utilized neural network to produce high-resolution bounding boxes. Researchers in [59] split the cross attention into branches with some focusing on classification and other on bounding box regression, aiming to increase object detection performance. Additionally, Authors in [60] innovatively employed unsupervised learning for object detection.

B. DISTANCE ESTIMATION

Distance estimation techniques in computer vision are versatile methods that can be applied across various sectors, including automotive and crime analysis. Numerous researchers proposed their unique methods for distance estimation, contributing to a rich body of research in this field. The research [61] proposed a method for integrating distance estimation within the YOLOv3 model. The authors updated YOLOv3 by incorporating a distance prediction vector and introduced a dedicated distance estimation loss function. This enabled the model to effectively learn and utilize distance information during training. In a similar vein, authors in [62] also undertook a comparable approach, but with a focus on depth estimation. In their experiments, they augmented the YOLOv4 network by introducing an additional depth estimation output channel branch. This branch was trained using 3D box labels enriched with depth information, and a dedicated depth loss function was employed to associate the model's output channel with this depth information. As a result, their model achieved impressive results, boasting an Average Precision (AP) of 71.68% for cars and an AP of 62.12% for pedestrians when evaluated on the KITTI dataset [63].

In the controversy, a few researchers have put forward innovative ideas that do not necessarily rely on datasets enriched with distance information. Authors in [64] took a distinctive approach by utilizing a system with two horizontally separated cameras. Their method was rooted in the principles of stereoscopy, and it allowed for distance calculation based on parameters such as the distance between the two cameras and the disparities between horizontal pixel values. As their proposed method primarily leveraged stereoscopic principles, it reduced the need for datasets specifically enhanced with distance information.

Authors in [65] introduced a different approach that harnessed the synergy between radar information and image data to ensure precise and reliable distance estimation. Their proposed method employed a middle-fusion technique, which combined radar point clouds and RGB images to enhance object detection and distance estimation accuracy in autonomous driving scenarios. In this innovative process, both radar and image data were independently responsible for estimating the distance to objects. The key aspect of their approach was the fusion of the results generated by these two methods. Importantly, during their experiments, the authors observed that radar outperformed image data in terms of distance estimation. Consequently, they implemented a mechanism to override redundant results, giving preference to more accurate radar-based estimates when necessary.

Currently, various studies on object detection have demonstrated the robustness of models in terms of lightweight design. However, distance estimation methods often rely on radar data, which require a lot of information. In this research, we propose a model that can reduce the information needed by using 2D images and achieve better object detection performance compared to existing models.

III. METHODOLOGY

In this research, we present a YOLOv8-CAW model that incorporates both the Coordinate Attention (CA) module and the Wise Intersection over Union (WIoU) loss function. Additionally, this research introduces a novel distance estimation algorithm to extend the capabilities of the YOLOv8-CAW model. The structure of Section III is organized as follows: Section (A) provides an in-depth discussion of the original YOLOv8 model, Section (B) explores the integration of the CA module with the YOLOv8-CAW model, Section (C) elaborates on the WIoU loss function, and Section (D) delves into the details of the distance estimation algorithm.

A. YOLOv8 MODEL

Yolov8 is the latest release of the YOLO family series. The first YOLO model was developed by Redmon et al. [66]. The YOLO model series has always been famous because of its superiority in terms of object detection; after a few iterations, the YOLOv8 model was introduced by Jocher et al. [4]. YOLOv8 builds upon the success of the previous YOLO series by introducing several key improvements. One significant change is the adoption of an anchor-free detection head. Unlike previous versions that relied on anchor boxes, YOLOv8 directly predicts the object's center, eliminating the need of IOU matching or assigning scales on one side and chooses a task-aligned method to match positive and negative samples. This simplifies the model and enhances its ability to handle small or overlapping objects. Besides that, YOLOv8 replaces the C3 module (CSPDarknet53 with 3 convolutions) from YOLOv5 with a novel C2f module. The C2f module reduces the network by one convolutional layer based on the original C3 module. This module enhances computational speed without compromising performance. It achieves this by effectively combining the strengths of past YOLO models. The C2f module facilitates a richer flow of gradient information while maintaining a lightweight structure, as illustrated in Figure 1.

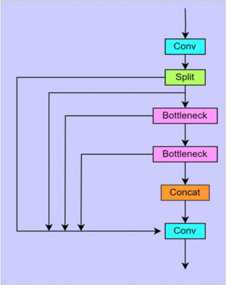


FIGURE 1. Architecture of C2F module.

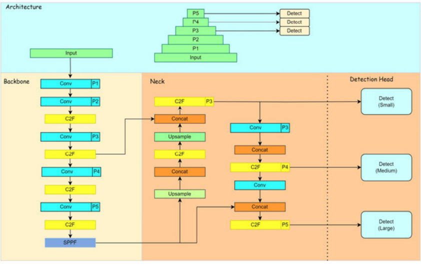


FIGURE 2. Original YOLOv8 Architecture *Referred and modified based on Rangeking's illustration on GitHub [67].

Figure 2 illustrates the overall YOLOv8 architecture. Inspired by the Path Aggregation Network (PAN) [68], YOLOv8 merges features extracted from various resolution feature maps to generate multi-scale features. The backbone takes the input image and down samples it five times, resulting in five scale feature maps denoted as {*P1, P2, P3, P4, P5*}. The primary function of the backbone is to extract these informative features from the input image. Following the backbone is the neck, which acts as a bridge between the extracted features and the detection head. The neck refines these features and facilitates their fusion, ensuring the detection head can leverage the most informative representation for object detection. Finally, the detection head will take responsibility for object detection tasks. It is noteworthy to mention that, in YOLOv8, the scale set {*P3, P4, P5*} plays a crucial role, with each scale specifically responsible for detecting objects of a particular size range – small objects for P3, medium objects for P4, and large objects for P5. This division of labor across scales strengthens the model's overall detection accuracy.

B. COORDINATE ATTENTION MODULE

The effectiveness of attention mechanisms had been demonstrated in the era of rapid computer science development, for example, research focused on the utilization of channel attention, as explored in [69], or the mixed attention proposed by Jiang et al. [70], has demonstrated the remarkable efficacy of attention mechanisms. Among the various attention mechanisms, there are also some standout mechanisms like the

Convolutional Block Attention Module developed (CBAM) by Woo et al. [35] or Squeeze-and-Excitation Networks (SE) developed by Hu et al. [10] that have been repeatedly experimented with and demonstrated their superiority. After reviewing the existing attention mechanisms, the CA module has been chosen in this research because CA is a novel attention mechanism that embeds positional information into channel attention, where the network can focus on large important regions at little computational cost. CA has also been shown to achieve state-of-the-art results on a variety of computer vision tasks, including image classification, object detection, and semantic segmentation. The recent research in [18] and [71], also demonstrated CA has the potential to surpass SE and CBAM in the overall performance.

As demonstrated in [10], the SE block could be divided into squeeze and excitation steps. Given the input X, the squeeze step for the c-th channel can be formulated as the equation (1).

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{1}$$

where X $= [x_1, x_2, \ldots, x_c]$ is the intermedia tensor and the z_c is the output associated with the c-th channel. SE uses global pooling to encode global spatial information, compressing the global information into a scalar, which makes it difficult to retain important spatial information. Because of this, CA converted global pooling into two 1-dimensional encoding operations by modifying equation (1), the new equation can be formulated as shown in equations (2) and (3).

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leq i \leq W} x_c(h, i) \tag{2}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leq i \leq H} x_c(j, w) \tag{3}$$

The modified formula integrates features from different directions to output a pair of direction-aware feature maps. Compared with the compression method of global pooling, this allows the attention block to capture long-distance relationships in a single direction while preserving spatial information in another direction, helping the network to locate targets more accurately.

After obtaining the 2-direction pooling output, the map of attention is generated by concatenating the outputs from equations (2) and (3) into a shared 1 × 1 convolutional transformation function $F1$, the process can be formulated in equation (4).

$$\mathbf{f} = \delta(F_1([z^h, z^w])) \tag{4}$$

where $[z^h, z^w]$ means that the spatial dimension has been concatenated, and δ is a non-linear activation function. After obtaining non-linear data through the activation function, the output is then divided into two groups of feature maps according to the horizontal and vertical directions.

$$\mathbf{f^h} \in \mathbb{R}^{C/r \times H} \tag{5}$$

$$\mathbf{f^w} \in \mathbb{R}^{C/r \times W} \tag{6}$$

Where r denotes the reduction ratio for controlling the block size. The two 1 × 1 convolution transforms are represented by F_h, F_w and σ function is employed to transform the output from equations (5) and (6) to tensors with the same channel number as the input X, the formulation is shown in equations (7) and (8).

$$\mathbf{g}^h = \sigma(F_h(\mathbf{f^h})) \tag{7}$$

$$\mathbf{g^w} = \sigma(F_w(\mathbf{f^w})) \tag{8}$$

CA reduces the channel number of $\mathbf{f}$ with an appropriate r to reduce the overhead model complexity. In the end, output from equations (7) and (8) are employed as the attention weights, and the equation of the generation of attention block is formulated in equation (9).

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{9}$$

C. YOLOv8-CAW MODEL WITH COORDINATE ATTENTION MODULE

The CA module helps the model learn the regions in the feature map that are related to the target locations, thus improving the localization accuracy of the model. This research chose to place the CA module after the C2f module in YOLOv8 because the C2f module first decomposes the input feature map into channels and then compresses the features of each channel through convolutional layers and activation functions. Attention modules also help the model learn more important features from the compressed feature maps with minimized computational complexity, thus improving the detection performance of the model. The complete architecture is shown in Figure 3.

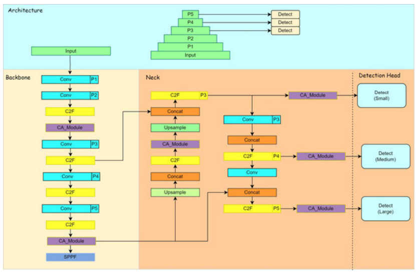


FIGURE 3. YOLOv8-CAW architecture with CA module.

D. WISE-IoU LOSS FUNCTION

The bounding box regression loss function is an essential component of object detection. The Intersection over Union (IoU) loss function for bounding box prediction is first introduced by Yu et al. [72]. During the training process, the model generates predicted boxes and ground truth boxes. The difference between these boxes, measured by IoU, indicates the model's accuracy, and the equation of IoU in [7] can be expressed as equation (10).

$$\text{IoU} = 1 - \frac{W_i H_i}{S_u} \tag{10}$$

The bounding box regression loss function is also used to measure the difference in position between the predicted box and the ground truth box and to optimize the model to adjust the position of the predicted box, thereby improving the model's accuracy. Many loss functions have already been developed and experimented with repeatedly by researchers, such as Distance-IoU (DIoU) and Complete-IoU (CIoU) by Zheng et al. [6] and Generalized-IoU (GIoU) by Rezatofighi et al. [73].

Among all these algorithms, this research chooses WIoU based on the one that requires a lightly accurate bounding box to support the distance estimation task. Apart from that, other researchers also shown that WIOU outperforms CIOU in their experiments [20], [40]. Compared with the traditional IoU method, WIoU has a 2 layers attention mechanism to react according to different quality of training data which can be expressed as:

$$L_{IoU} = 1 - IoU \tag{11}$$

$$R_{WIoU} = \exp(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}) \tag{12}$$

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{13}$$

An anchor box is defined by its center coordinates and size B =[$xywh$]. The ground truth box is defined by B_{gt} = [$x_{gt}y_{gt}w_{gt}h_{gt}$]. In equations (11) to (13), WIoU reduces the impact of the geometry factor, while the term amplifies the penalty from low-quality anchor boxes and decreases the penalty from high-quality anchor boxes. WIoU also focuses on the distance between the center points of the anchor box and the target when the two boxes overlap.

According to the inventor of WIoU [7], a dynamic non-monotonic focusing mechanism is used in WIoU. If compared to traditional IoU loss function algorithms, WIoU can dynamically adjust the loss coefficient for each training sample. For easy samples, the loss coefficient is reduced, which reduces the focus on these samples. For difficult samples, the loss coefficient is increased, which increases the focus on these samples. This helps the model to better learn difficult samples, which improves the final detection performance. WIoU adopts an outlier β as the vector to determine a low or high-quality anchor box, β is higher, then the smaller gradient gain will be assigned, to avoid low-quality data ruining the overall training performance. The β can be defined as:

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \in [0, +\infty] \tag{14}$$

To associate with the outlier calculation strategy, WIoU constructs a non-monotonic focusing coefficient r using β to build WIoUv3 which will be used in this research and can be expressed in equations (15) and (16).

$$r = \frac{\beta}{\delta \alpha^{\beta - \delta}} \tag{15}$$

$$L_{WIoUv3} = r L_{WIoUv1} \tag{16}$$

E. DISTANCE ESTIMATION

Distance estimation is the process of determining the distance between two points or objects. It is a fundamental task in many fields, including navigation, robotics, and computer vision. To measure distance using computer vision techniques, this research proposes a formula to calculate the ratio between the object size in real life and the object size in the computer vision view, which can be written as follows:

$$D_{istance} = \frac{O_{bject}S_{ize} \times F_{ocal}L_{ength}}{B_{ounding\ Box}S_{ize}} \tag{17}$$

$$F_{ocal}L_{ength} = \frac{D_{istance} \times B_{ounding\ Box}S_{ize}}{O_{bject}S_{ize}} \tag{18}$$

Fundamentally, the proposed distance estimation algorithm utilizes the principle in calculating the ratio between real-life object size and the size indicated by bounding boxes in object detection. Equation (17) outlines the process, starting with informing the computer of a specific object's size. Once the object is detected, the computer outlines it with a bounding box and computes the distance by comparing the given object size with the detected bounding box size, ensuring precise distance measurements. The focal length plays a crucial role in acquainting the computer with the camera lens's focus distance, minimizing the risk of miscalculations.

For Equation (18) concerning focal length calculation, both the object size and bounding box dimensions can be obtained through human comprehension and object detection, respectively. We demonstrate how backpropagation can be utilized to compute the focal length, enabling distance estimation even when the focal length is not initially provided. Pseudo code for manipulating the distance estimation program is also provided for better understanding.

Algorithm 1 Distance Estimation Algorithm

```
INPUT:
FL ← Camera Focal Length
Pred_Class ← Predicted Classes in Image
OS ← Object Size in Real World
BS ← Size of Object Bounding Box in Computer Vision View
OUTPUT:
D ← Estimated Distance
START:
Load source from Image
IF Pred_Class is predicted:
    Convert bounding box coordinates to BS
    D ← (FL x OS) / BS
ELSE
    Return Error
ENDIF
```

It's important to acknowledge that, in the current stage of development, if an object is too small to be discerned by the human eye at a certain distance, the proposed model may encounter challenges in handling such objects. This limitation arises from the focal length setting of our proposed model, which currently only accommodates original-sized images rather than enlarged ones. Consequently, the model may struggle to detect small objects.

IV. EXPERIMENT IMPLEMENTATION

A. YOLOv8-CAW TRAINING DATASET

The datasets used in this YOLOv8-CAW are both PASCAL VOC 2007 and PASCAL VOC 2012 datasets. The PASCAL VOC dataset consists of 20 different categories which include: person, bird, cat, cow, horse, sheep, airplane, bike, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, and TV monitor. Some of the examples of PASCAL VOC dataset images can be found in Figure 4.

FIGURE 4. Example Image of PASCAL VOC Dataset. The image codes for the photos from top left to bottom right are as follows: 00012, 000193, 001759 00050, 000131, 000109.

B. MODEL TRAINING

The experiments in this research were conducted on a system running the Windows 11 operating system, equipped with an Intel®Core™i9-12900H Processor and an NVIDIA GeForce RTX 3080 Ti Laptop Graphics Processing Unit (GPU) featuring 16 GB of graphics memory. The programming language employed for this research is Python, facilitating seamless implementation of various deep learning algorithms. Furthermore, the CUDA®parallel computing platform has been installed to harness the computational power of the GPU, thereby accelerating model training and experimentation processes.

TABLE 1. Initial parameters for model training.

Parameter	Setting
Input Size	640
Batch Size	32
Epochs	200
Initial Learning Rate	0.01
IoU Threshold	0.5
Optimizer	Auto

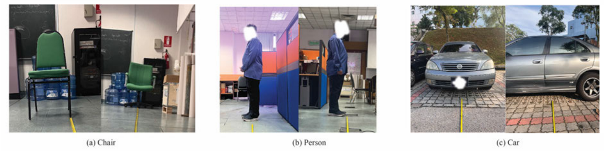
Input size refers to the dimensions of the images fed into the deep learning model, while batch size denotes the number of samples processed simultaneously during training. A larger batch size can expedite convergence and training speed but also necessitates more hardware memory. Epochs represent the total number of iterations required for completing model training, while the initial learning rate dictates the magnitude of weight updates during training. A higher learning rate may cause the model to overshoot optimal convergence, whereas a lower rate can lead to slower convergence. The IoU (Intersection over Union) threshold defines the degree of overlap between ground truth boxes and predicted boxes.

It's noteworthy that in the 'auto' optimizer strategy of YOLOv8, the model initially utilizes the AdamW optimizer for the first 10,000 iterations before transitioning to Stochastic Gradient Descent (SGD). This approach capitalizes on AdamW's ability to facilitate rapid convergence, particularly in the early training stages. However, as iterations progress, AdamW's efficacy may diminish, prompting a switch to the more sustainable and stable SGD optimizer after 10,000 iterations.

C. DISTANCE ESTIMATION TEST SAMPLES COLLECTION

To validate the proposed distance estimation algorithm, a benchmark experiment. In this experiment, the YOLOv8-CAW model is integrated with selected three distinct classes from the PASCAL VOC dataset to represent a range of object scales, which included small, medium, and large objects.

Later, the images of objects are captured by ranging the distance of 1, 2, and 3 meters in the laboratory, Multiple images are acquired for each object at different angles and with different objects. Figure 5 shows some examples of the test samples.



(a) Chair

(b) Person

(c) Car

FIGURE 5. Sample Test Instances for the Distance Estimation Benchmark Experiment, (a) Chair, (b) Person, (C) Car.

V. EXPERIMENT RESULTS

A. YOLOv8 BENCHMARK EXPERIMENT

To verify the proposed YOLOv8-CAW model, the recall, precision, mAP 0.5, 0.75, 0.5:0.95, and inference time are chosen as comparative metrics. The formula for calculating recall, precision, and mAP can be found in equation (19) to (22).

$$Recall(R) = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{19}$$

Recall refers to the percentage of samples that are correctly predicted out of all total positive samples.

$$Precision(P) = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{20}$$

Precision refers to the percentage of samples that are correctly predicted out of all samples that are predicted as positive. To calculate mAP, AP needs to be found first, and then mAP is calculated by taking the average of AP values for all classes, where the equation for both can be found in equations (21) and (22).

$$AP = \int_{0}^{1} Precision\,(Recall)\,d(Recall) \tag{21}$$

AP is equal to the area under the precision–receall curve, it gets higher when both precision and recall are high.

$$mAP = \frac{1}{N}\sum\nolimits_{i=1}^{N} AP_i \tag{22}$$

The metric mAP measures the AP across N classes. mAP at an IoU (Intersection over Union) threshold of 0.5, referred to as mAP@0.5, evaluates precision when the IoU threshold is set to 0.5. The same principle applies to mAP@0.75. mAP@0.5:0.95, on the other hand, calculates the mAP across a range of IoU thresholds, from 0.5 to 0.95. In the realm of object detection, the mAP is consistently recognized as a pivotal evaluation metric, indicative of the model's performance and accuracy. Additionally, this paper will also adopt model parameters, GFLOPs (Giga Floating Point Operations Per Second), and inference time (IT) as the model computational complexity metric.

This research introduces a novel YOLOv8-CAW model that incorporates the CA module and employs the Wise-IoU (WIoU) loss function. The study includes a comprehensive comparison involving various model configurations, such as the YOLOv8 baseline model with and without WIoU, an improved YOLOv8 model that combines the CA module with and without WIoU, and the YOLOv8 model with the CA module but without the WIoU loss function. The outcomes of these comparisons are presented in Table 2. To make Table 2 more understandable, a few abbreviations have been made, and 'W' indicates WIoU Loss Function. 'CB' Indicates the CBAM module, and 'CA' Indicates the Coordinate Attention Module.

TABLE 2. Model comparison.

Method	Params (Million)	GFLOPs	R	P	mAP@ 0.5	mAP@ 0.75	mAP@ 0.5:0.95	IT (ms)
YOlOv8	25.85	78.7	77.6	82.0	84.2	71.6	65.6	12.8
YOLOv8+W	25.85	78.7	77.4	83.7	84.8	72.0	66.8	**12.3**
YOLOv8+CB	26.86	79.1	77.0	83.5	84.6	71.8	65.7	14.9
YOLOv8+CB+W	26.86	79.1	**78.3**	82.2	85.3	72.7	66.4	14.6
YOLOv8+CA	26.05	79.6	78.2	82.6	85.5	73.2	67.0	13.4
YOLOv8-CAW	26.06	79.6	78.1	**84.2**	**85.6**	**73.5**	**67.1**	12.6

Table 2 demonstrates that our proposed model delivers the highest performance in terms of mAP. It is noteworthy that our proposed model exhibits a slightly lower recall percentage when compared to the combination of CBAM with WIoU. However, the CBAM combination results in lower precision, leading to a reduced mAP. Additionally, the proposed model has effectively maintained a low inference time speed compared to the baseline model. While the baseline model employing WIoU loss function achieved the lowest inference time speed, YOLOv8-CAW demonstrated even lower inference time speed compared to the plain baseline model. However, the integration of CBAM led to the slowest inference time speed. This observation suggests that the YOLOv8-CAW model could substantially enhance model performance while still ensuring a low inference time speed.

To have a better observation of the model, a heatmap is presented using Eigen-Cam [74]. The heatmaps can highlight specific regions of feature maps on which different models concentrate as shown in Figure 6.

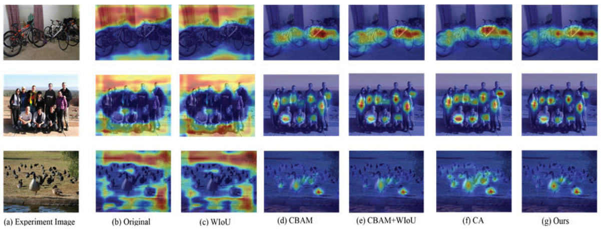


FIGURE 6. Visualized heatmap, (a) Experiment Image, (b) Original YOLOv8 model, (c) YOLOv8 with WIoU, (d) YOLOv8 with CBAM, (e) YOLOv8 with CBAM & WIoU, (f) YOLOv8 with CA, (g) Our proposed model. The image codes that come from the PASCAL VOC dataset in (a) from top to bottom are as follows: 2009_003175, 005205, 001275.

By analyzing Figure 6, the following can be observed. Firstly, the original YOLOv8 model appears to have poor attention to the objects in question. In contrast, when the YOLOv8 model is integrated with the Attention Module, it demonstrates a substantial improvement in feature map attention. Additionally, by comparing Figure 6(f) and Figure 6(d), we can observe that the CA module appears to offer more effective feature map attention than CBAM.

To establish the superiority of our proposed model over existing researchers, we adopted a dual validation approach. Initially, comparisons were conducted with recent researchers to ensure objectivity. Subsequently, different researchers on YOLOv8 model were selected for further validation, employing identical hyperparameters as provided by the respective authors. Additionally, experiments were conducted on both the PASCAL VOC dataset and the larger MS-COCO dataset to demonstrate the model's performance across varying scales of datasets. Dataset sizes are provided in Table 3, while the results of unidentical hyperparameter model comparisons are shown in Table 4 and Table 5. While the identical hyperparameter results will be provided in Tables 7 and 8. The hyperparameter settings as provided by the authors are detailed in Table 6.

TABLE 3. Dataset attributes.

Dataset	No. of Class	Training Images	Validation Set
VOC	20	16551	4952
COCO	80	118287	5000

Table 3 provides details on the dataset size and distribution for both training and validation sets. To ensure an objective comparison of results.

Table 4 compares the performance of various object detection methods on the PASCAL VOC dataset. Proposed YOLOv8-CAW demonstrates superior performance with a mAP of 85.6%, surpassing other recent models. This underscores its efficacy in object detection. However, it is important to note that while our proposed model achieves the highest

TABLE 4. Models comparison on pascal Voc test 20007 dataset.

Method	Input Size	Params (Million)	GFLOPs	mAP@ 0.5
YOLO-Anti [11]	416 x 416	-	-	85.5
MFFAMM [23]	300 x 300	-	-	80.7
LKC-Net [26]	640 x 640	7.28	-	84.0
Faster R-CNN w/ CEFPN [28]	1000 x 600	-	-	81.3
YOLO-Former [33]	640 x 640	26.13	13.8	83.0
Mini-YOLOv4-tiny [49]	288 x 288	3.79	-	**72.07**
DPNet [50]	320 x 320	2.5	1.0	81.5
Improved YOLOv5 [51]	640 x 640	3.3	7.7	79.3
YOLOv4-EEEA-Net-C2 [55]	-	31.15	5.54	81.8
Bagging R-CNN [57]	1000 x 600	-	-	81.4
YOLOv8-CAW (Ours)	640 x640	26.06	79.6	**85.6**

TABLE 5. Models comparison on MS-COCO Val2007 dataset.

Method	Input Size	Params (Million)	AP@ 0.5:0.95	AP@0.5	AP@0.75
EYOLOX [19]	640 x 640	13.54	42.2	61.6	45.4
Faster R-CNN+AFPN [27]	800 x 1000	52.2	41.9	61.3	45.4
CF-YOLO [39]	-	22.0	36.1	55.8	-
Trident-YOLO [48]	416 x 416	-	18.8	37.0	17.3
LNFCOS [51]	800 x 1333	27.1	37.2	56.0	39.9
YOLO-ERF-S [52]	640 x 640	5.9	41.3	60.7	-
FastDARTSDet [56]	640 x 640	6.9	-	59.4	41.7
LEOD-Net [58]	560 x 560	-	**48.11**	53.21	44.33
DESTR-DC5-R101 [59]	-	88.0	46.4	**67.1**	50.1
DETReg [60]	-	-	45.5	64.1	49.9
YOLOv8-CAW (Ours)	640 x 640	26.1	47.2	64.2	**51.4**

mAP among all models, several models exhibit comparable performance. For instance, YOLO-Anti achieves a mAP of 85.5%, trailing our proposed model by only 0.1%. Similarly, LKC-Net achieves a mAP of 84.0%, ranking as the third best among all models. Despite the narrow margins, our models consistently demonstrate the best overall performance, showcasing significant potential in the field of object detection.

Table 5 presents a comparison of object detection methods on the MS-COCO dataset. It is noteworthy that our proposed model achieved the second-best result, with a 47.2% AP across IoU thresholds from 0.5 to 0.95, the LEOD-Net model outperformed with a 48.11% AP, claiming the top position. However, our proposed model demonstrated superiority in AP at IoU thresholds of 0.5 and 0.75, achieving 64.2% and 51.4%, respectively, compared to LEOD-Net's 53.21% and 44.33%. Similarly, in AP at an IoU threshold of 0.5, our proposed model attained the second-best result at 64.2%, while the DESTR-DC5-R101 model achieved 67.1% as the best result. Despite this, our proposed model showcased better performance across IoU thresholds from 0.5 to 0.95 and an IoU threshold of 0.75, with respective APs of 47.2% and 51.4%, whereas DESTR-DC5-R101 achieved 46.4% and 50.1%, respectively. In conclusion, by observing the overall performance of different models, the proposed model significantly demonstrates superior performance compared to other models on the MS-COCO dataset.

TABLE 6. Hyperparameter setting provided by original authors.

Method	Input Size	Batch Size	Epochs	I.Lr	IoU Threshold	Optimizer	Dataset
YOLOv8-CAB [22]	640	32	300	0.001	0.2	AUTO	COCO
Enhanced YOLOv8 [42]	640	32	300	0.01	0.7	SGD	VOC
YOLOv8-CGRNet [45]	640	16	200	0.01	0.5	AUTO	VOC
PGDS-YOLOv8s [53]	640	32	100	0.01	0.5	AUTO	COCO

Table 6 presents the hyperparameter settings provided by the original authors I.Lr indicates the initial learning rate in short form. It is worth noting that the YOLOv8 model was configured to halt training once the model's performance remained unchanged for the last 50 epochs. Hence, for references in [29] and [37]. Initially, both comparative experiments were intended to run for 300 epochs each. However, due to our model's performance stop evolving from the last 50 epochs, training automatically ceased at epochs 259 and 157, respectively. The comparative results of the PASCAL VOC dataset are presented in Table 7 and the MS-COCO dataset is presented in Table 8.

TABLE 7. Identical hyperparameter comparison on pascal Voc dataset.

Method	Params (Million)	mAP@0.5
Enhanced YOLOv8 [42]	8.5	84.4
YOLOv8-CAW (Ours)	26.06	**85.8**
YOLOv8-CGRNet [45]	-	81.9
YOLOv8-CAW (Ours)	26.06	**85.7**

TABLE 8. Identical hyperparameter comparison on MS-COCO dataset.

Method	Params (Million)	AP@ 0.5:0.95	AP@0.5	AP@0.75	mAP@ 0.5	mAP@ 0.5:0.95
YOLOv8-CAB [22]	-	-	-	-	47.1	28.2
YOLOv8-CAW (Ours)	26.1	-	-	-	**61.8**	**45.4**
PGDS-YOLOv8s [53]	10.81	43.5	60.1	47.1	-	-
YOLOv8-CAW (Ours)	26.1	**47.2**	**64.2**	**51.4**	-	-

Table 7 displays a comparison of recent research using identical hyperparameters. Upon careful examination of Table 7, it's evident that the proposed model consistently achieves higher performance compared to other models. Therefore, based on the results obtained from the PASCAL VOC dataset, we confidently conclude that the proposed model demonstrates superior performance compared to other models.

Table 8 presents a comparison of other models using identical hyperparameters with the MS-COCO dataset. In the first model's comparison, the proposed model demonstrates comprehensive superiority in performance. Similarly, the second model's comparison also exhibits a comprehensive improvement. Therefore, we can conclude that even when utilizing a larger dataset, the proposed model maintains competitive performance and shows great potential. Despite multiple comparisons having been made and verified, to have a clearer understanding of the proposed model, the Precision-Recall (PR) graph and confusion matrix for both the PASCAL VOC dataset and MS-COCO dataset are also provided in Figure 7 for better understanding.

A PR recall curve graph illustrates the relationship between precision and recall for the model. Precision measures the accuracy of positive predictions, while recall describes the ability of the model to capture all positive samples. The curve demonstrates how changing the threshold affects these metrics. Ideally, the curve that is closer to the upper-right corner indicates high precision and recall, which means the model performs better. As for the confusion matrix, we need to consider the numerous categories used in this research, which can make the information presented in the images unclear. Therefore, we should approach it from a different perspective. In a confusion matrix, the x-axis represents the real samples, while the y-axis corresponds to the predicted samples. Ideally, higher values on both axes indicate better model performance. Subsequently, in the provided confusion matrix, darker colors indicate higher values.

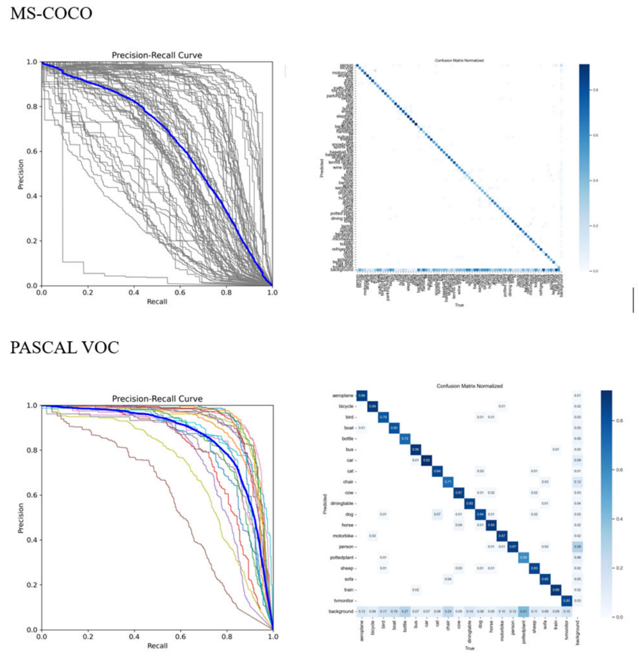


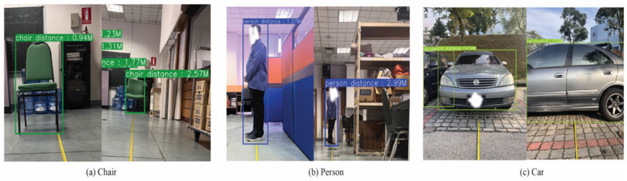
FIGURE 7. Precision-Recall (PR) Curve and Confusion Matrix for both PASCAL VOC dataset and MS-COCO dataset.

Consequently, we will describe the confusion matrix by examining the distribution of colors. In Figure 7, the PR curve on the MS-COCO dataset doesn't exhibit perfect performance. The presence of a mid-positioned line suggests the model's performance is mediocre. However, considering the dataset's large number of samples and categories, the results remain acceptable. In the confusion matrix, a relatively clear diagonal line is visible in the middle, indicating that the proposed model makes relatively accurate predictions across all categories. On the other hand, the PR curve on the PASCAL VOC dataset demonstrates superior performance compared to the MS-COCO dataset. The curve on the PASCAL VOC dataset appears closer to the right-upper corner and begins to form a distinct shape, indicating the model's excellent performance. Performance in confusion matrix for both datasets, a clear diagonal line is observed in the middle, signifying that the proposed model makes relatively accurate predictions across all categories.

B. DISTANCE ESTIMATION RESULT VALIDATION

Following the categorization of objects by class and distance, each category will include 10 images captured at distances of 1, 2, and 3 meters in the laboratory. A distance estimation experiment is conducted at the laboratory on all the captured images and computes the average value for each category to establish a global reference. Sample images showcasing the distance estimation process are depicted in Figure 8, and all detected distances are recorded in the list from Table 9 to 11, which represents 1, 2, and 3 meters accordingly.

Based on the records in Table 9, the overall algorithm performance is deemed acceptable. However, there are still noticeable deviations, such as the minimum value for "car" and the maximum value for "person." Nevertheless, the overall average deviation remains within 0.1 meters. Regarding Table 10, while the performance of 'Car' has significantly declined, the rest of the categories still maintain high performance. In the controversy concerning long distances, the overall model performance has significantly deteriorated. Even though the prediction accuracy is high, it can be observed that the maximum value for 'Person' is only 0.01m away from the ground truth value, and the average performance has also experienced a substantial decline. Hence, it can be concluded that the algorithm's performance decreases as the distance increases. Nonetheless, the proposed algorithm still demonstrates its potential capability to handle distance estimation tasks, as the overall average detection accuracy remains approximately 90% close to the ground truth distance, except for the 'Car' class in long-distance detection.



(a) Chair

(b) Person

(c) Car

FIGURE 8. Detected sample test instances for the distance estimation benchmark experiment, (a) Chair, (b) Person, (C) Car.

TABLE 9. Detected distance list in 1 meter.

Category	1 M	Average	Min	Max
Chair	[0.94,1.07,0.85,1.10,1.03,1.09,0.85,1.05,0.97,1.11]	**1.01**	**0.85**	**1.11**
Person	[0.98,1.18,1.07,1.04,1.09,1.02,0.96,1.00,1.12,1.16]	**1.06**	**0.96**	**1.18**
Car	[0.74,1.03,0.86,0.94,0.84,1.09,0.89,1.07,1.03,1.07]	**0.96**	**0.74**	**1.09**

TABLE 10. Detected distance list in 2 meter.

Category	2 M	Average	Min	Max
Chair	[1.94,2.05,2.11,1.88,2.00,2.09,1.82,2.14,1.97,2.03]	**1.98**	**1.82**	**2.14**
Person	[1.92,1.95,2.01,1.89,1.86,2.03,2.00,1.88,1.98,1.87]	**1.93**	**1.86**	**2.03**
Car	[1.84,1.92,1.73,1.91,1.79,1.88,1.75,1.70,1.93,1.76]	**1.83**	**1.70**	**1.93**

TABLE 11. Detected distance list in 3 meter.

Category	3 M	Average	Min	Max
Chair	[2.57,2.63,2.51,2.74,2.85,2.79,2.57,2.60,2.91,2.78]	**2.70**	**2.51**	**2.91**
Person	[2.99,2.75,2.84,2.95,2.89,2.90,2.79,2.81,2.77,2.91]	**2.86**	**2.75**	**2.99**
Car	[2.75,2.63,2.79,2.58,2.67,2.52,2.82,2.61,2.70,2.76]	**2.68**	**2.52**	**2.82**

VI. CONCLUSION

In this research, a YOLOv8-CAW model, and a distance estimation algorithm based on pure computer vision methods are introduced. Model complexity is always an issue in deep learning development, and the CIoU loss function that binds with the original YOLOv8 model has a significant limitation. Therefore, the YOLOv8-CAW model is proposed, augmented with a lightweight attention mechanism that significantly boosts performance while adding minimal parameters to the model. Additionally, the new WIouU loss function performs better than the original loss function. While certain evaluation metrics may indicate superiority for other models, our proposed model maintains low inference times while significantly improving the model's mAP. Simultaneously, we compared our proposed model with others using various approaches and datasets. This comprehensive analysis highlights the superiority of our model, demonstrating its significant potential.

Furthermore, a distance estimation technique integrated with the YOLOv8-CAW model which eliminates the need for additional information and only requires a single 2D image is also proposed. In experiments for distance estimation, although algorithm performance decreases with increasing distance, the proposed algorithm still exhibits significant potential for handling distance estimation tasks. This is attributed to the fact that, despite some deviations, most average detections remain approximately above 90% close to the ground truth distance. Meanwhile, one of the applications of computer vision-based distance estimation is in the autonomous driving industry, where it can provide better accuracy in detecting and assessing distances between vehicles or objects.

The current proposed model is encountering limitations in terms of generalization. The performance of our approach exhibits a correlation with object size, even within the same category. For example, variations in vehicle models can lead to different results. Therefore, a finer-grained classification dataset is essential to address these challenges. As a suggested future endeavor, expanding the dataset would be considered. Additionally, the optimization of the distance estimation algorithm to ensure consistent performance at various distances should also be interesting.

ACKNOWLEDGMENT

The authors extend their gratitude to the entire Multimedia University community for their valuable support during the distance estimation benchmark experiment. Special thanks are also extended to all researchers who contributed their techniques and datasets that are adopted in this research.

REFERENCES

[1] N. A. Othman, M. U. Salur, M. Karakose, and I. Aydin, "An embedded real-time object detection and measurement of its size," in *Proc. Int. Conf. Artif. Intell. Data Process. (IDAP)*, Malatya, Turkey, 2018, pp. 1–4, doi: 10.1109/IDAP.2018.8620812.

[2] J. R. Martinez-de Dios, B. C. Arrue, A. Ollero, L. Merino, and F. Gómez-Rodríguez, "Computer vision techniques for forest fire perception," *Image Vis. Comput.*, vol. 26, no. 4, pp. 550–562, Apr. 2008, doi: 10.1016/j.imavis.2007.07.002.

[3] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. 14th Int. Conf.*, 2011, pp. 332–339.

[4] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *YOLO By Ultralytics (Version 8.0.0)*. [Online]. Available: https://github.com/ultralytics/ultralytics

[5] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," 2021, *arXiv:2103.02907*.

[6] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," 2019, *arXiv:1911.08287*.

[7] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," 2023, *arXiv:2301.10051*.

[8] C. Jiang, H. Zhang, Y. Yue, and X. Hu, "AM-YOLO: Improved YOLOV4 based on attention mechanism and multi-feature fusion," in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, vol. 6, Mar. 2022, pp. 1403–1407, doi: 10.1109/ITOEC53115.2022.9734536.

[9] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9197–9206.

[10] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*.

[11] K. Wang and M. Liu, "YOLO-Anti: YOLO-based counterattack model for unseen congested object detection," *Pattern Recognit.*, vol. 131, Nov. 2022, Art. no. 108814, doi: 10.1016/j.patcog.2022.108814.

[12] W. Yang, D. Bo, and L. S. Tong, "TS-YOLO: An efficient YOLO network for multi-scale object detection," in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, vol. 6, Mar. 2022, pp. 656–660, doi: 10.1109/ITOEC53115.2022.9734458.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[14] H. Lou, X. Duan, J. Guo, H. Liu, J. Gu, L. Bi, and H. Chen, "DC-YOLOV8: Small-size object detection algorithm based on camera sensor," *Electronics*, vol. 12, no. 10, p. 2323, May 2023, doi: 10.3390/electronics12102323.

[15] Y. Chen, H. Liu, J. Chen, J. Hu, and E. Zheng, "Insu-YOLO: An insulator defect detection algorithm based on multiscale feature fusion," *Electronics*, vol. 12, no. 15, p. 3210, Jul. 2023, doi: 10.3390/electronics12153210.

[16] H. Li, J. Li, H. Wei, Z. Liu, Z. Zhan, and Q. Ren, "Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles," 2022, *arXiv:2206.02424*.

[17] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Sep. 2019, pp. 3007–3016.

[18] W. Xuan, G. Jian-She, H. Bo-Jie, W. Zong-Shan, D. Hong-Wei, and W. Jie, "A lightweight modified YOLOX network using coordinate attention mechanism for PCB surface defect detection," *IEEE Sensors J.*, vol. 22, no. 21, pp. 20910–20920, Nov. 2022, doi: 10.1109/JSEN.2022.3208580. https://doi.org/10.1109/JSEN.2022.3208580

[19] R. Tang, H. Sun, D. Liu, H. Xu, M. Qi, and J. Kong, "EYOLOX: An efficient one-stage object detection network based on YOLOX," *Appl. Sci.*, vol. 13, no. 3, p. 1506, Jan. 2023, doi: 10.3390/app13031506.

[20] G. Wang, Y. Chen, P. An, H. Hong, J. Hu, and T. Huang, "UAV-YOLOV8: A small-object-detection model based on improved YOLOV8 for UAV aerial photography scenarios," *Sensors*, vol. 23, no. 16, p. 7190, Aug. 2023, doi: 10.3390/s23167190.

[21] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2023, pp. 10323–10333.

[22] M. Talib, A. H. Y. Al-Noori, and J. Suad, "YOLOV8-CAB: Improved YOLOV8 for real-time object detection," *Karbala Int. J. Modern Sci.*, vol. 10, no. 1, pp. 56–68, Jan. 2024, doi: 10.33640/2405-609x.3339.

[23] Z. Qu, T. Han, and T. Yi, "MFFAMM: A small object detection with multi-scale feature fusion and attention mechanism module," *Appl. Sci.*, vol. 12, no. 18, p. 8940, Sep. 2022, doi: 10.3390/app12188940.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf.*, 2015, pp. 21–37.

[25] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," 2017, *arXiv:1711.07767*.

[26] W. Wang, S. Li, J. Shao, and H. Jumahong, "LKC-Net: Large kernel convolution object detection network," *Sci. Rep.*, vol. 13, no. 1, p. 9535, Jun. 2023, doi: 10.1038/s41598-023-36724-x.

[27] G. Yang, J. Lei, Z. Zhu, S. Cheng, Z. Feng, and R. Liang, "AFPN: Asymptotic feature pyramid network for object detection," in *Proc. IEEE Int. Conf. Syst. Man, Cybern. (SMC)*, Oct. 2023, pp. 2184–2189, doi: 10.1109/SMC53992.2023.10394415.

[28] Y. Luo, X. Cao, J. Zhang, J. Guo, H. Shen, T. Wang, and Q. Feng, "CE-FPN: Enhancing channel information for object detection," *Multimedia Tools Appl.*, vol. 81, no. 21, pp. 30685–30704, Sep. 2022, doi: 10.1007/s11042-022-11940-1.

[29] Y. Dai and W. Liu, "GL-YOLO-lite: A novel lightweight fallen person detection model," *Entropy*, vol. 25, no. 4, p. 587, Mar. 2023, doi: 10.3390/e25040587.

[30] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, Feb. 2023.

[31] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," 2021, *arXiv:2101.03697*.

[32] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," 2018, *arXiv:1804.06882*.

[33] Y. Dai, W. Liu, H. Wang, W. Xie, and K. Long, "YOLO-Former: Marrying YOLO and transformer for foreign object detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022, doi: 10.1109/TIM.2022.3219468. https://doi.org/10.1109/TIM.2022.3219468

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[35] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[36] L. Zhang, G. Ding, C. Li, and D. Li, "DCF-YOLOV8: An improved algorithm for aggregating low-level features to detect agricultural pests and diseases," *Agronomy*, vol. 13, no. 8, p. 2012, Jul. 2023, doi: 10.3390/agronomy13082012.

[37] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," 2017, *arXiv:1702.03118*.

[38] D. Misra, "Mish: A self regularized non-monotonic activation function," 2019, *arXiv:1908.08681*.

[39] Q. Ding, P. Li, X. Yan, D. Shi, L. Liang, W. Wang, H. Xie, J. Li, and M. Wei, "CF-YOLO: Cross fusion YOLO for object detection in adverse weather with a high-quality real snow dataset," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 1–11, Jul. 2023, doi: 10.1109/TITS.2023.3285035.

[40] Q. Zhao, H. Wei, and X. Zhai, "Improving tire specification character recognition in the YOLOV5 network," *Appl. Sci.*, vol. 13, no. 12, p. 7310, Jun. 2023, doi: 10.3390/app13127310.

[41] G. Liu, Y. Hu, Z. Chen, J. Guo, and P. Ni, "Lightweight object detection algorithm for robots with improved YOLOV5," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106217, doi: 10.1016/j.engappai.2023.106217.

[42] M. Zhang, Z. Wang, W. Song, D. Zhao, and H. Zhao, "Efficient small-object detection in underwater images using the enhanced YOLOV8 network," *Appl. Sci.*, vol. 14, no. 3, p. 1095, Jan. 2024, doi: 10.3390/app14031095.

[43] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run don't walk: Chasing higher FLOPS for faster neural networks," 2023, *arXiv:2303.03667*.

[44] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4.

[45] Y. Niu, W. Cheng, C. Shi, and S. Fan, "YOLOV8-CGRNet: A lightweight object detection network leveraging context guidance and deep residual learning," *Electronics*, vol. 13, no. 1, p. 43, Dec. 2023, doi: 10.3390/electronics13010043.

[46] T. Wu, S. Tang, R. Zhang, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," 2018, *arXiv:1811.08201*.

[47] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758. https://doi.org/10.1109/TPAMI.2019.2938758

[48] G. Wang, H. Ding, B. Li, R. Nie, and Y. Zhao, "Trident-YOLO: Improving the precision and speed of mobile device object detection," *IET Image Process.*, vol. 16, no. 1, pp. 145–157, Jan. 2022, doi: 10.1049/ipr2.12340.

[49] D.-S. Bacea and F. Oniga, "Single stage architecture for improved accuracy real-time object detection on mobile devices," *Image Vis. Comput.*, vol. 130, Feb. 2023, Art. no. 104613, doi: 10.1016/j.imavis.2022.104613.

[50] Q. Zhou, H. Shi, W. Xiang, B. Kang, X. Wu, and L. Jan Latecki, "DPNet: Dual-path network for real-time object detection with lightweight attention," 2022, *arXiv:2209.13933*.

[51] B. Hwang, S. Lee, and H. Han, "LNFCOS: Efficient object detection through deep learning based on LNblock," *Electronics*, vol. 11, no. 17, p. 2783, Sep. 2022, doi: 10.3390/electronics11172783.

[52] X. Wang, N. He, C. Hong, F. Sun, W. Han, and Q. Wang, "YOLO-ERF: Lightweight object detector for UAV aerial images," *Multimedia Syst.*, vol. 29, no. 6, pp. 3329–3339, Dec. 2023, doi: 10.1007/s00530-023-01182-y.

[53] D. Yang, J. Zhou, T. Song, X. Zhang, and Y. Song, "PGDS-YOLOV8S: An improved YOLOV8S model for object detection in fisheye images," *Appl. Sci.*, vol. 14, no. 1, p. 44, Dec. 2023, doi: 10.3390/app14010044.

[54] 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, *arXiv:1405.0312*.

[55] C. Termritthikun, Y. Jamtsho, J. Ieamsaard, P. Muneesawang, and I. Lee, "EEEA-Net: An early exit evolutionary neural architecture search," *Eng. Appl. Artif. Intell.*, vol. 104, Sep. 2021, Art. no. 104397, doi: 10.1016/j.engappai.2021.104397.

[56] C. Wang, X. Wang, Y. Wang, S. Hu, H. Chen, X. Gu, J. Yan, and T. He, "FastDARTSDet: Fast differentiable architecture joint search on backbone and FPN for object detection," *Appl. Sci.*, vol. 12, no. 20, p. 10530, Oct. 2022, doi: 10.3390/app122010530.

[57] P. Li, Y. He, D. Yin, F. R. Yu, and P. Song, "Bagging R-CNN: Ensemble for object detection in complex traffic scenes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10097085.

[58] H. Ibrahem, A. Salem, and H.-S. Kang, "LEOD-Net: Learning line-encoded bounding boxes for real-time object detection," *Sensors*, vol. 22, no. 10, p. 3699, May 2022, doi: 10.3390/s22103699.

[59] L. He and S. Todorovic, "DESTR: Object detection with split transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9377–9386.

[60] A Bar, X Wang, V Kantorov, C. J. Reed, R Herzig, G Chechik, A Rohrbach, T Darrell, and A. Globerson, "DETReg: Unsupervised pretraining with region priors for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 14605–14615.

[61] M. Vajgl, P. Hurtik, and T. Nejezchleba, "Dist-YOLO: Fast object detection with distance estimation," *Appl. Sci.*, vol. 12, no. 3, p. 1354, Jan. 2022, doi: 10.3390/app12031354.

[62] J. Yu and H. Choi, "YOLO MDE: Object detection with monocular depth estimation," *Electronics*, vol. 11, no. 1, p. 76, 2022.

[63] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 3354–3361, doi: 10.1109/CVPR.2012.6248074.

[64] B. Strbac, M. Gostovic, Z. Lukac, and D. Samardzija, "YOLO multi-camera object detection and distance estimation," in *Proc. Zooming Innov. Consum. Technol. Conf. (ZINC)*, May 2020, pp. 26–30, doi: 10.1109/ZINC50678.2020.9161805.

[65] R. Nabati and H. Qi, "Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles," 2020, *arXiv:2009.08428*.

[66] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[67] RangeKing. (2023). *Brief Summary of YOLOv8 Model Structure*. [Online]. Available: https://github.com/ultralytics/ultralytics/issues/189

[68] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2018, pp. 8759–8768.

[69] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023, doi: 10.1016/j.inffus.2023.03.021.

[70] K. Jiang, Z. Wang, P. Yi, C. Chen, Z. Han, T. Lu, B. Huang, and J. Jiang, "Decomposition makes better rain removal: An improved attention-guided deraining network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3981–3995, Oct. 2021, doi: 10.1109/TCSVT.2020.3044887.

[71] X. Li, B. Wu, X. Zhu, and H. Yang, "Consecutively missing seismic data interpolation based on coordinate attention unet," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3128511. https://doi.org/10.1109/LGRS.2021.3128511

[72] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 516–520, doi: 10.1145/2964284.2967274.

[73] H. Rezatofighi, N. Tsoi, J. Y. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.

[74] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7, doi: 10.1109/IJCNN48605.2020.9206626.

ZU JUN KHOW received the B.Comp.Sc. degree (Hons.) in artificial intelligence from Multimedia University, Malaysia, in 2022, where he is currently pursuing the M.Eng.Sc. degree. His research interests include image processing and deep learning.

YI-FEI TAN received the B.Sc. (Hons.), M.Sc., and Ph.D. degrees from the University of Malaya, Malaysia. She is currently an Associate Professor with the Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia. Her research interests include machine learning, deep learning, image processing, big data analytics, and queueing theory.

HEZERUL ABDUL KARIM (Senior Member, IEEE) received the B.Eng. degree in electronics with communications from the University of Wales Swansea, U.K., in 1998, the M.Eng. degree in science from Multimedia University, Malaysia, in 2003, and the Ph.D. degree from the University of Surrey, U.K., in 2008. He is currently a Professor with the Faculty of Engineering, Multimedia University. His research interests include telemetry, error resilience and multiple description video coding for 2D/3D image/video coding and transmission, and content-based image/video recognition. He serving as the Chair for the IEEE Signal Processing Society Malaysia Chapter.

HAIRUL AZHAR ABDUL RASHID (Senior Member, IEEE) received the B.Eng. degree in electronic and electrical engineering from University College London, U.K., in 1997, and the M.Eng.Sc. degree in signal processing and the Ph.D. degree in optical communication systems from Multimedia University, Malaysia, in 2001 and 2007, respectively. He is currently a Professor with the Faculty of Engineering, Multimedia University. His research interest includes specialty optical fiber for sensing applications.