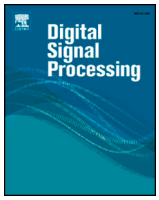
Digit. Signal Process. 155 (2024) 104765

Contents lists available at ScienceDirect

Digital Signal Processing

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



THDet: A Lightweight and Efficient Traffic Helmet Object Detector based on YOLOv8

Yi Li [a], Huiying Xu [a,*], Xinzhong Zhu [a,b,d], Xiao Huang [c], Hongbo Li [d]

[a] School of Computer Science and Technology of Zhejiang Normal University, Jinhua, Zhejiang, 321004, China
[b] Research Institute of Hangzhou Artificial Intelligence, Zhejiang Normal University, Hangzhou, Zhejiang, 311231, China
[c] College of Education of Zhejiang Normal University, Jinhua, Zhejiang, 321004, China
[d] Beijing Geekplus Technology Co., Ltd, Beijing, 100101, China

ARTICLE INFO

Keywords:
Object Detection
Traffic Helmet Detection
Feature Fusion
Lightweight Network

ABSTRACT

Traffic helmet object detection is playing an increasing important role in the smart traffic fields. However, object size variation and small-shaped helmet detection has still been a challenging problem by reason of their poor visual appearance in the image. In this work, we present an efficient traffic helmet detector through feature enhancement and lightweight design based on YOLOv8n called THDet. Specifically, we employ the coordinate attention into C2f blocks combined with softmax activate function to achieve feature channel aggregation and strong non-linear expression of the backbone for further effective feature extraction; Next, Focal_CIoU loss function embedded with Focal Loss method is utilized for the more precise measure of various objects bounding box regression and balance of positive and negative examples during training; Then, a new lightweight detection head style is designed only with two proper position heads (P3 & P4) to perform final classification and localization, through this scheme saving the 33.7% parameters than baseline method. Finally, Attention Refined Features Module (ARFM) is built to calibrate the multi-scale fused features by introducing 3-D weights generated from SimAttention to boost the final detection accuracy. Extensive experiments have demonstrated that our proposed method realizes noticeable performance in terms of detection accuracy and inference speed compared with baseline YOLOv8n and many end-to-end detectors of similar model size. Concretely, THDet achieves 0.447 at the overall evaluation metric of $mAP_{0.5-0.95}$, accomplishing 3.2% detection accuracy improvement than YOLOv8n. Besides, THDet only holds 2.2M parameters with 295 FPS inference speed, reducing 33.4% parameters compared with YOLOv8n. The experimental results validate the effectiveness of our proposed method, showcasing that THDet outperforms the mainstream real-time detection algorithms in the terms of accuracy, inference speed and lightweight model design for traffic helmet object detection.

1. Introduction

Traffic helmet object detection have received increasing attention in the recent years with the growing number of vehicles in the road. How to correctly wear a helmet and protective clothes properly is a key to avoid the occurrence of traffic accidents, minimizing the unnecessary loss while protecting personal life. At the present, the common way to recognize whether the rider wears a helmet usually supervised by the policemen who in charge of directing traffic at the main crossroads. This kind of method not only causes excessive human resources but also leads to low efficiency in the more densely populated sites. Therefore, raising people's awareness of wearing helmets and improving the efficiency of helmet detection and identification is currently a hot issue in the field of traffic safety that needs to be solved urgently.

As one of the most crucial and challenging tasks in the computer vision, object detection has achieved remarkable progress with the advance of deep learning and have been applied successfully in the miscellaneous application scenarios [1], such as automotive driving, UAV, video monitor, robot video and etc. Its mission involves identifying and localizing objects in an digital image to understand what and where the target is in the predefined classes, providing valuable information for advanced semantic understanding of images and videos. Object de-

* Corresponding author.
E-mail addresses: leeyee@zjnu.edu.cn (Y. Li), xhy@zjnu.edu.cn (H. Xu), zxz@zjnu.edu.cn (X. Zhu), huangxiao@zjnu.edu.cn (X. Huang), Jason.li@geekplus.com (H. Li).

https://doi.org/10.1016/j.dsp.2024.104765

Available online 7 September 2024

tection performance have been largely improved due to the constant evolution of deep learning technology. Deep convolutional neural network (DCNN) for feature extraction has been widely used in computer vision tasks [2] due to its strong representation ability for abstract features. Meanwhile, object detection has completed the transformation from the traditional hand-designed feature based detection method to the deep learning DCNN methodology. Then the object detection algorithms based on convolutional neural network (CNN) rapidly became the mainstream research in the field of image processing. Generally, there exist two mainstream object detection categories, the one is two-stage detectors, the most representative is RCNN [3] series. The other is one-stage detectors, such as YOLO [4] series. Two-stage detectors owe the advantage in high localization and object recognition accuracy, while one-stage detectors are more time-efficient when applied in the industry for real-time object detection. Two-stage object detectors usually follow two process paradigm to perform target detection: the first step generates candidate region proposals and then using the features extracted from the former step to perform target classification and localization by detector.

The current object detection algorithms tend to perform well for large and medium objects while achieve unpleasant detection results on small targets due to the limited available features and various shapes presented in the image [5]. Small object detection is still a challenging study in computer vision field, insufficient visual information hampers the extraction of discriminative features for classification and localization when detecting small objects. In the traffic helmet scene, the difficulty of object detection is further aggravated by the multiple overlap objects scale and the weak texture features of small targets. Therefore, directly introducing an extra high-resolution pyramid level brings more noise in the bottom of the feature fusion stage. These mentioned issues significantly lower the quality of helmet features and have a negative impact on the helmet detection.

Benefiting from renovation of deep neural networks of modeling ability for scale variation and abstraction for feature information, performance of small object detection have received unprecedented growth. Many researches aim to improve the performance of this task in the various approaches, including constructing more powerful feature extraction backbone [6] [7], designing effective feature fusion methodology [8] [9], embedding channel and spatial attention mechanism more semantic information for different size objects [10], design new label assignment strategies for better samples distribution [11] [12] etc. However, the improvement of general small object detection is usually focused on the common dataset like COCO [5], for the application scenario like traffic safety field still need further study to reach high detection performance especially in the complex traffic environment.

Deep learning object detectors have been constantly evolving with needs of multitudes application, each having its own set of requirements. Specifically, some safety-critical applications require high accuracy and low-latency, which emphasizes the importance of computation-saving and energy-efficient networks. Real-time detection has always been critical issue when its implementation on the real-world. Many deep learning object detectors seek the trade-off between the accuracy and inference speed. A wide range of studies endeavor to speed up the inference process like design lightweight backbone network with less parameters and complexity [13] [14]. Among the numerous object detection algorithms, YOLO (***You Look Only Once***) series detection framework become spotlight for its excellent balance of speed and accuracy. In the field of traffic helmet detection tasks, it needs heavily not only high detection accuracy when coming to mass targets gathered in a focused region, but also realizes the competent real-time online detection which help the model to deploy on the edge devices with limited computing and memory resources.

To achieve excellent detection performance with high accuracy and high inference speed for traffic helmet detection task, we propose a real-time detector called THDet, which involves overall lightweight backbone design, enhanced feature extraction pattern with attention, optimization for Bounding Box Regression (BBR) loss function, lightweight choices for detection heads selection and recalibration of multi-scaled features for final detection heads. THDet improves the detection 3.2% accuracy compared with baseline, meeting the real-time detection at the evaluation standards of 295 FPS with 0.447 $mAP@_{0.50-0.95}$. We demonstrate the effectiveness of our framework by conducting experiments on the *Traffic Helmet Dataset*. Experiments show that THDet has strong ability when applied on the traffic helmet detection, which enables us to train detectors that significantly outperforms other mainstream object detectors regarding on the different evaluation metrics. The main contributions of this work are as follows:

- We propose a lightweight, efficient framework for traffic helmet detection named THDet, which is combined with high effective feature extraction block, lightweight detection heads, 3-D weights refinement for feature fusion and productive bounding box regression loss function, achieving the real-time standard with high detection accuracy.
- We applied Coordinate Attention (CA) into the C2f blocks combined with softplus activation function by embedding positional information into channel dimension to capture long range dependencies of complex features; Focal_CIoU loss function was adopted to realize fast and accurate bounding box regression with balanced samples.
- To realize the lightweight model paradigm, we have reduced the detection heads from 3 to 2, which saving large margin parameters of the model and speeding up the inference process notably; to mitigate the accuracy loss from the reduced heads, we construct Attention Refined Features Module (ARFM) consists of the SimAttention which directly generates 3-D weights (both spatial and channel) for considering the importance of difference sizes of objects in the detection head.
- Extensive experiments have been conducted to verify the effectiveness of our proposed THDet, combining the above mentioned improvement metrics, our method achieved detection performance 0.447 at $mAP@_{0.5-0.95}$ with 295 FPS and saved 33.4% parameters compared with the baseline YOLOv8n. Besides, THDet outperforms the mainstream object detectors on the various evaluation standards and realizes balanced trade-off between accuracy and inference speed, meeting the needs of the real-time detection in the filed traffic safety detection scenario.

2. Related works

2.1. Deep Learning Object Detectors

Great progress has been made for the object detection in the recent years thanks to the booming development of DCNN, which transformed this study from traditional hand-crafted design to data driven detectors. Deep learning based object detection methods usually can be classified into two categories: two-stage detectors and one-stage detectors. RCNN is the first two-stage milestone deep learning detector proposed by Girshick et al., which demonstrates how CNN can be used to immensely improve performance on the PASCAL VOC 2007 [15] dataset, ushering a new wave in the field object detection. Faster RCNN [16] introduced a fully convolutional network called Region Proposal Networks (RPN) that takes arbitrary input image which largely speed up and simplify the complicated region generation step of RCNN.

YOLO series are receiving increasing popularity when deploy the object detection in the real applications. As one of the most representative one-stage algorithms, YOLO regards the detection as a regression pattern, the whole network architecture treats the input features for only one time, which avoids the time-consuming part for generating region proposals and achieves high inference speed. Currently, the YOLO algorithm has been evolved to YOLOv10 [17] with more overall improvements from different perspectives to further improve the detection performance. Multi-scale feature object detection is introduced by SSD

[18], with this ingenious design, small object detection has been largely improved to a new level. On the other hand, SSD neglects the interaction of different level features, which limits the overall detection performance when image information gets complicate. There are still many SSD variants to perfect this kind of detection methodology such as RFBDet [19] and M2Det [20].

Transformer is a kind of newly proposed neural network which utilize self-attention mechanism to extract intrinsic feature, which is originally applied in natural language processing filed and brings in significant improvement in computer vision. The Transformer-based detectors (DETRS) [21] have received extensive attention from academia since it was proposed due to its elimination of various hand-crafted components, like non-maximum attention (NMS). This kind of network design simplifies the pipeline of object detection a large margin and realizes end-to-end object detection. RT-DETR [22] proposed an efficient hybrid encoder consists of intra-scale interaction and cross-scale feature fusion module, accelerating the divergence of training process and reaching end-to-end object detection.

2.2. Object Detection in Traffic Field

There exists abundance of research works when coming to the traffic field object detection. Wang et al. [23] proposed an improved YOLOv5 network applied to traffic sign detection by designing an improved feature pyramid fusion process, which combines attention mechanism and feature reuse techniques to reduce the information loss and boost the feature pyramid presentation. Lian et al. [24] introduced a small object detection method in traffic scenes by integrating multi-scale channel attention block, more effective contextual information block of feature fusion, reaching excellent performance on the public dataset. Zeng et al. [25] introduced an efficient traffic sign detection method by combining transformer-style architecture and a residual classifier loss to improve the small size traffic sign detection performance. Lee et al. [26] proposed YOLO-EfficientNet to improve the helmet detection performance by splitting the training process into two part, one stage for detection head, another is for categorizing them into multiple classes. To address low accuracy and slow detection of helmet detection of YOLOv4 model, Li et al. [27] proposed a lightweight helmet detector called YOLO-PL, by designing YOLO-P algorithm to improve the small object detection and label assignment. Besides, lightweight VoVNet was introduced to realize effective feature extraction with cost-effective parameters. Mi et al. [28] designed an intelligent helmet detection model based YOLOv5, by combining pixel level attention and coordinate attention together to build advanced feature extraction blocks and also applying CPAG-Net as the detection network. The result indicates the proposed method performs efficiently with higher detection accuracy for helmet detection.

3. THDet: Improved YOLOv8n Model for Traffic Helmet Object Detection

In this section, We will demonstrate the proposed THDet and improved modules in detail. First of all, we will show the whole architecture of THDet in Section 3.1. Second, the enhanced feature extraction of C2f block with CA and softplus activation function is presented in Section 3.2. Next, the optimization for BBR loss function with Focal_CIoU is introduced in Section 3.3. Then, to achieve lightweight network paradigm and ameliorate the computation burden of model parameter and complexity, we develop the optimization detection heads skills in Section 3.4. Finally, a type of refined features with 3-D weights combined with SimAttention mechanism is built for fast and accurate detection in Section 3.5.

3.1. THDet Architecture

YOLO attracts increasing attention for its superior performance in terms of detection accuracy and inference speed with the high confidence to classify and locate the target in digital image. The YOLO series algorithms have evolved multiple versions to perfect its detection performance. YOLOv8 [29] is released by *ultralytics* in Jan. 2023, drawing many latest excellent object detection strategies together, in attempt to find the best trade-off between detection accuracy and inference performance.

In this work, we adopt the prevailing currently one-stage object detector YOLOv8n as our baseline by virtue of its excellent trade-off between detection accuracy and inference speed. YOLOv8 algorithm involves 5 categories for various detection scenario according to different scales of model depth and width, including n, s, m, l and x respectively. The difference of the above classes' parameters and model complexity is increasing with better performance as well as the running speed descends. We choose the smallest size model YOLOv8n as the baseline with acceptable parameters and competitive detection results, combined with effective improved metrics to construct THDet.

Fig. 1 shows the overall architecture of THDet, including four main parts: Input digital images; Backbone with 4 stages for feature extraction process, each stage is composed of C2f block with coordinate attention and softplus convolutions; Neck for multi-scale feature fusion with bottom-up and top-down information flows for concatenation; Finally, Head part is to realize the object classification and localization, which contains Attention Refined Features Module (ARFM) for heads with two scales size feature map 80×80, 40×40 to execute object detection task.

3.2. Enhanced Feature Extraction of C2f Block with CA and Softplus

Designing strong and robust backbone to extract sufficient feature is critical for computer vision downstream tasks. In recent years, remarkable progress has been made in the field of novel architecture design. Our work for improving the feature extraction backbone is not introducing new architecture but following inter-relation of original convolutional blocks to gradually perfect extraction process without pretrained weights on the large scale dataset like ImageNet [2].

THDet follows the YOLOv8 feature extraction C2f convolution blocks like Fig. 3 (b), the big difference between C3 3 (a) block from YOLOv5 and C2f block is the that latter one has more repeated bottlenecks and more split branches for feature concatenation, with this handcrafted design the backbone is capable to explore more detailed abstract information during CNN training in the early stages. Specially, we adopt the softplus activation function as non-linear transformation operation in our basic C2f convolution modules for its smooth characteristic in derivative near 0 point and preserve the non-linear property compared with ReLU. We have also conducted experiments to validate the effectiveness of different activation functions to convolutional blocks in Table 2. Fig. 2 list the visualization of compared activation functions.

To further strengthen the extraction ability of C2f modules, we have employ Coordinate Attention (CA) [30] incorporated with Bottleneck block in C2f, as shown Fig. 5. In the miscellaneous attention types, channel attention and spatial attention are the two representative mechanisms in the computer vision, which conduct weight recalibration and region selection respectively. CA is the combination of channel and spatial attention, it adaptively selects both important object and interest regions. Concretely, CA embeds positional information into channel attention, so that the long-range dependencies can be captured along one spatial direction, focusing on large important regions at little computational cost, as shown in Fig. 4. Channel features within CA are decomposed into two parallel one-dimensional model encodings, which contain directional distinguishable information, enabling each multi-scale feature map to acquire longer-range information along a spatial direction. Therefore, the model can more accurately locate and identify object areas.

The coordinate attention mechanism has two consecutive steps, coordinate information embedding and coordinate attention generation. First, two parallel average pooling branches encode the channel information horizontally and vertically. Next, a shared 1×1 convolutional layer is used to concatenate the outputs of the two pooling aggrega-

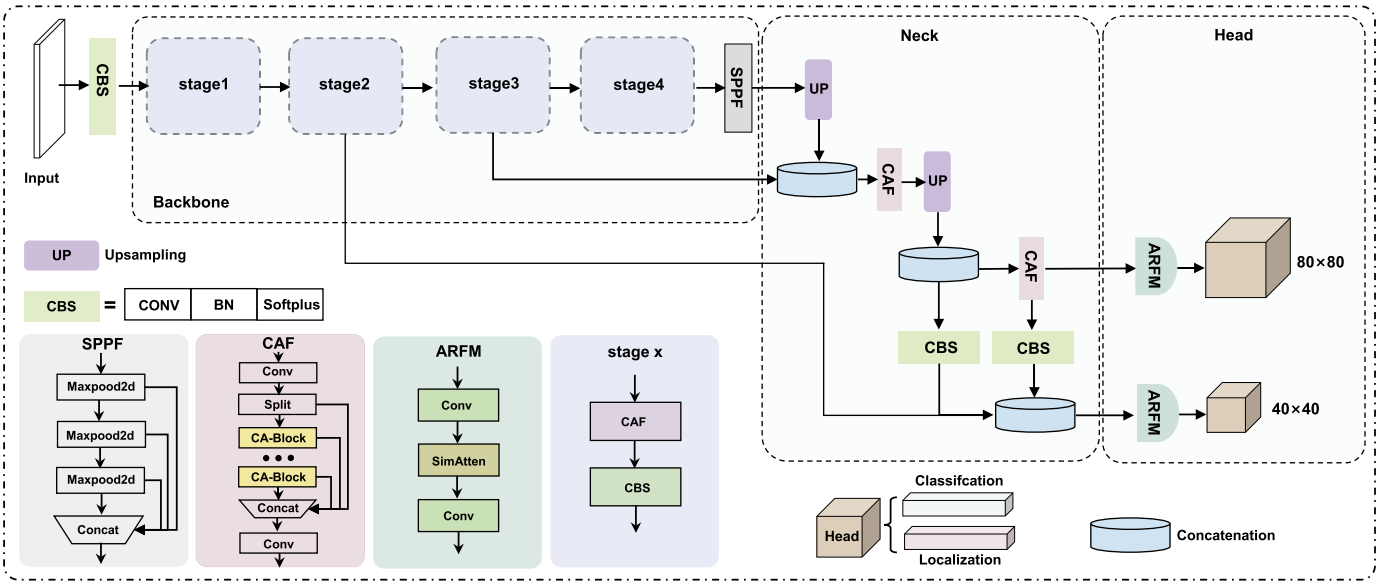


Fig. 1. Overall architecture of proposed THDet. Our method consists of four main components: input digital images; general backbone includes four stages with enhanced C2f name CAF and CBS block for effective feature extraction, SPPF with split branches maxpooling operation for advanced semantic aggregation; Neck is responsible for multi-scale feature fusion and concatenation, CAF module combined Coordinate Attention and C2f with several branches for further concatenation; Decoupled head is applied for classification and bounding box regression followed by ARFM, which uses SimAttention for refining fused multi-scales features.

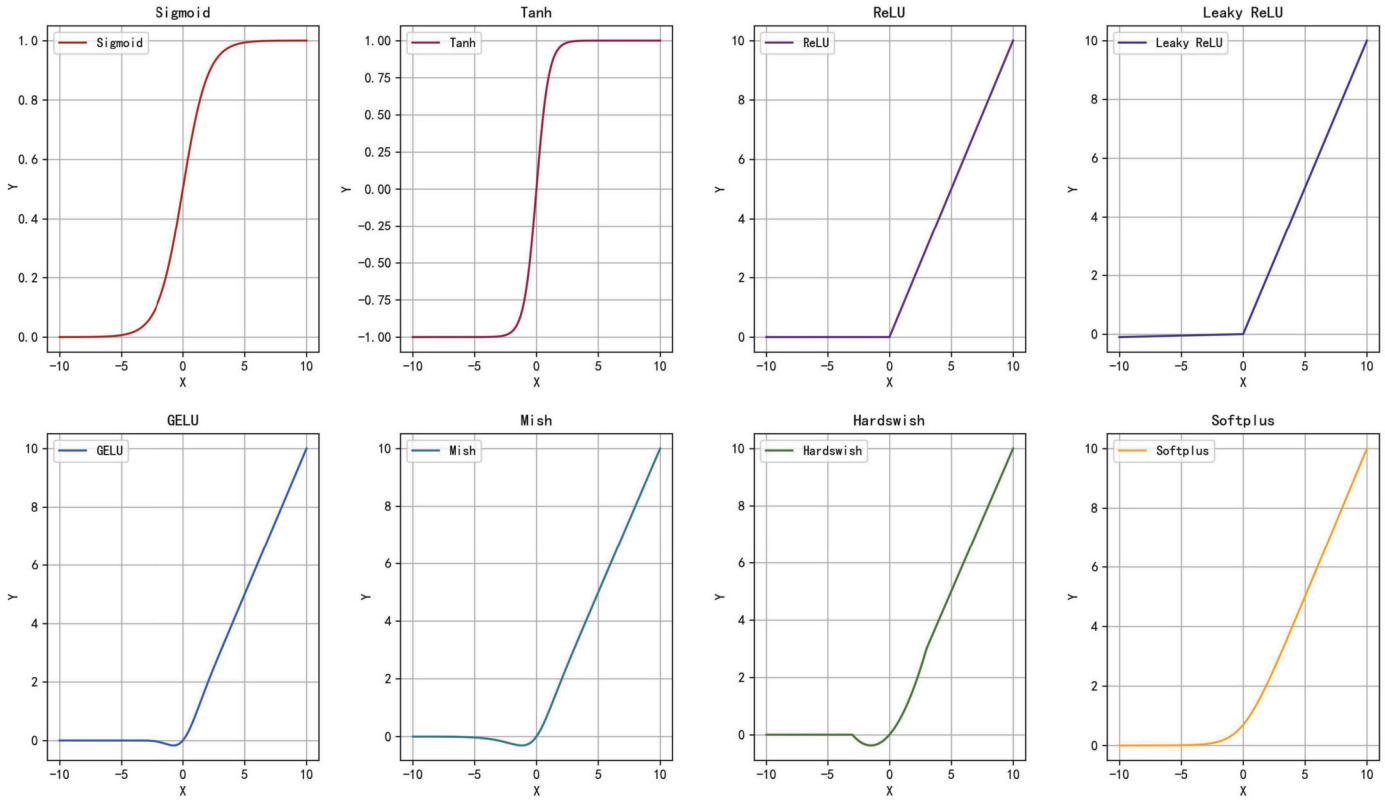


Fig. 2. Visualization of various activation functions in the convolutional blocks.

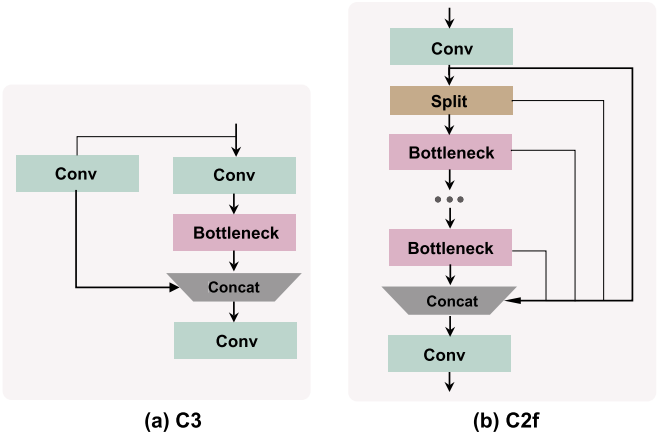


Fig. 3. The architecture C3 and C2f blocks for extracting features in the backbone.

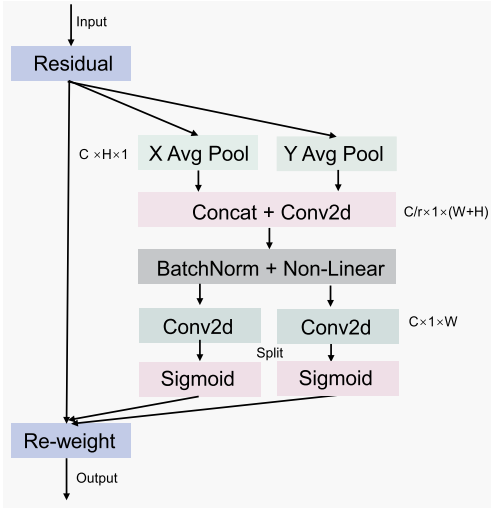


Fig. 4. The architecture of Coordinate Attention.

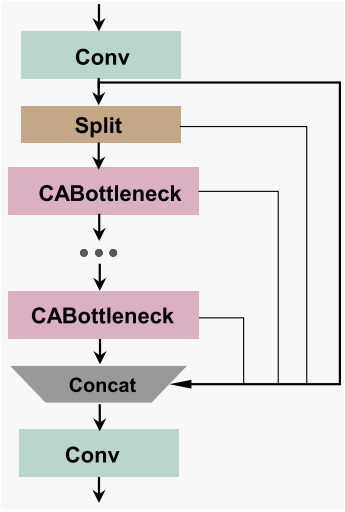


Fig. 5. Enhanced C2f blocks with Coordinate Attention.

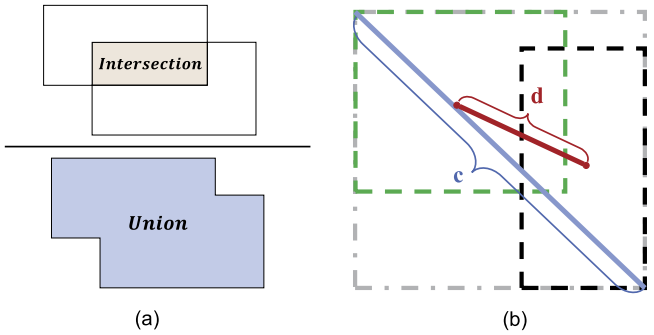


Fig. 6. (a) IoU calculation process, (b) CIoU, c is the diagonal length of the smallest enclosing box coverting two boxes, and $d = \rho(\mathbf{p}, \mathbf{p}^{gt})$ is the distance of central points of two boxes.

tion information. Then, coordinate attention divides the output results into two separate tensors by BatchNormalization and Non-linear activation function to yield attention vectors with the same number of channels for horizontal and vertical coordinates of the input X along. After that, the separate branch feature dimension w and h are process by general convolutional layer and sigmoid activation function to realize the smooth purpose for the output features. Finally, all the refined features are combined together along with the original residual inputs. The whole computation process can be represented as

$$f^h = GAP^h(X)$$
$$f^w = GAP^w(X) \tag{1}$$

$$d = \delta(BN(Conv_1^{1\times1}([f^h; f^w]))) \tag{2}$$
$$d^h, d^w = Split(d)$$

$$m^h = \sigma Conv_h^{1\times1}(d^h)$$
$$m^w = \sigma Conv_w^{1\times1}(d^w) \tag{3}$$

$$Out = Xm^h m^w \tag{4}$$

where GAP^h and GAP^w means pooling function and m^h and m^w denotes corresponding attention weights.

Employing coordinate attention embedded into C2f block, the network can accurately obtain the position of a targeted object. This approach enables the backbone has larger receptive field and also models cross-channel relationships at a lower cost, effectively enhancing the expressive power of the learned features.

3.3. Effective Localization Optimization for Bounding Box Regression

Loss function for Bounding Box Regression plays a critical role in object detection, which filters the excessive redundant proposals. Its good definition will bring significant performance improvement to the final detection results. $Intersection\ over\ Union$ is commonly used as object detection loss function metric in BBR, as shown in Fig. 6 (a) and Eq. (5), which calculates the ratio of intersection over union of predicted proposals and ground truth bounding box.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

where $A \cap B$ means the intersection between predicted proposals and ground truth and $A \cup B$ denotes the union of that equivalent.

Our THDet adopts the CIoU [31] as the BBR loss function as shown in 6 (b). Current loss methods still not sufficient to differentiate the hard regression examples during training and remain time-consuming to filter the redundant boxes during model inference. For example, GIoU [32] achieves competitive results than $\mathcal{L}_n$-norm methods, but GIoU still has limitation because of its insufficient of overlap area computation. CIoU takes overlap area, normalized central point distances and aspect ratio into account to regress overlap area scale with high efficiency.

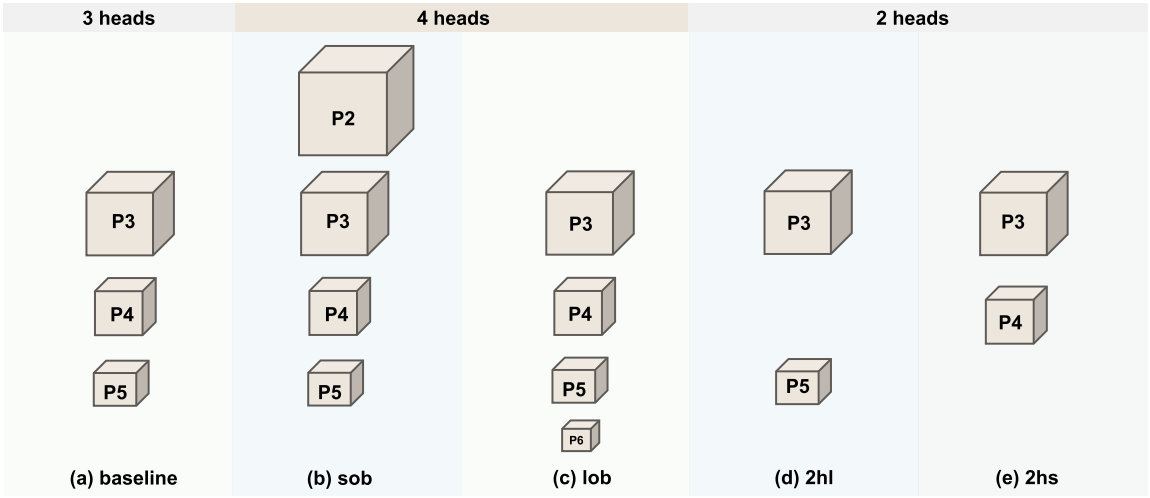


Fig. 7. Different methodologies for heads design. (a) baseline detection heads. (b) sob: adding extra P2 head for small object detection. (c) lob: adding extra P6 for large object detection. (d) 2hl: only two heads for large object detection: P3 and P5. (e) 2hs: only two heads for small object detection: P3 and P4.

$$GIoU = 1 - IoU + \frac{|\mathcal{A} - \mathcal{B} \cup \mathcal{B}^{gt}|}{|\mathcal{A}|} \tag{6}$$

where $\mathcal{A}$ is the smallest box area covering $\mathcal{B}$ and $\mathcal{B}^{gt}$,

CIoU considers the geometric factors for modeling regression process, which can be defined as,

$$\mathcal{L} = O(\mathcal{B}, \mathcal{B}^{gt}) + D(\mathcal{B}, \mathcal{B}^{gt}) + R(\mathcal{B}, \mathcal{B}^{gt}), \tag{7}$$

$$D = \frac{\rho^2(\mathbf{p}, \mathbf{p}^{gt})}{c^2} \tag{8}$$

where O, D, R denotes the overlap area, distance and aspect ratio respectively. $\mathbf{p}$ and $\mathbf{p}^{gt}$ are the central points of box $\mathcal{B}$ and $\mathcal{B}^{gt}$, c is the diagonal length of box $\mathbf{C}$ and ρ is the Euclidean distance coefficient.

$$M = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{9}$$

$$CIoU = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha M \tag{10}$$

where α is a trade-off parameter.

Imbalanced problems in training samples are inevitable in BBR, which mainly reflects on extreme inequality between the number of positive examples versus the number of negatives. If not addressed well, this imbalance will greatly impair detection accuracy. In this work, we adopt Focal Loss [33] as an optimized strategy to assist CIoU for fast training divergence, the final BBR loss function can be expressed,

$$\begin{aligned} Focal\_CIoU &= IoU^{\gamma} \times CIoU \\ &= (\frac{|A \cap B|}{|A \cup B| + \epsilon})^{\gamma} \times CIoU \end{aligned} \tag{11}$$

where γ is to measure inhibition of outliers.

3.4. Different Choices for Detection Lightweight head Design

Multi-scale features for prediction are generally accepted in the modern object detection algorithms since the introduction of Feature Pyramid Networks (FPN) which first formulates different sizes features by top-down and bottom-up information flows for fusion and concatenation and enriches the fused features with both semantic and contextual details, benefiting the overall detection performance.

We have designed several types of detection heads to validate the influences to model size, complexity and detection inference speed. According to traditional principles, more heads will do benefits to detection performance, Fig. 7 (a) shows the baseline detection heads modes with only three scale size 80×80, 40×40 and 20×20 from large to small

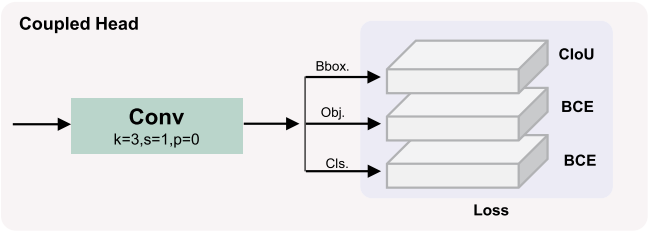


Fig. 8. Coupled head for detection.

respectively. Adding an extra 160×160 size P2 head with baseline for improving detection small target, as shown in (b). Similarly, adding an extra 10×10 size P6 head with baseline benefits performance on large target, as shown in (c). Counterintuitively, we also devise two heads for detection like (d) and (e) with P3 in common, but P4 and P5 was chosen for distinctive locations. We have fully conducted experiments in Table 5 to validate the influences of various methods from (a)-(e) to the object detection performance. According the experimental data shown, using 4 heads for detection cannot improve detection performance and will introduce vast of parameters to model like P6 head. However, by adopting 2 heads for detection, large parameters have been dropped with acceptable detection accuracy lost and remarkable inference speed growth. Considering the trade-off of model overall accuracy and running speed, we finally chose (e) mode for THDet.

3.5. Attention Refined Features Module for Efficient Detection Head

In object detection, the conflict between classification and regression task is another imbalanced problem, which limits the performance of the detector to some extent. The paradigm of decoupled head is well employed in the two-stage object detectors. To put it from another angle, decoupled head separates the detection task into two dependent pipelines, while coupled head treats this task at the same pipeline which will lead to inefficiency for training divergence, as is shown in Fig. 8 and Fig. 9. Our THDet follows the decoupled head method for object detection. Specifically, decoupled head first contains two parallel 1×1 branch to adjust channel number and then followed by two parallel 3×3 convolutions to execute the final classification and localization.

In order to build an efficient head for fast and accurate detection, a channel and spatial combined attention mechanism named SimAttention [34] as is shown in Fig. 10 (c), is utilized to construct Attention Refined Feature Module (ARFM). As we know, traditional convolution operation adopts static weight and lacks adaptability, which is insuffi-

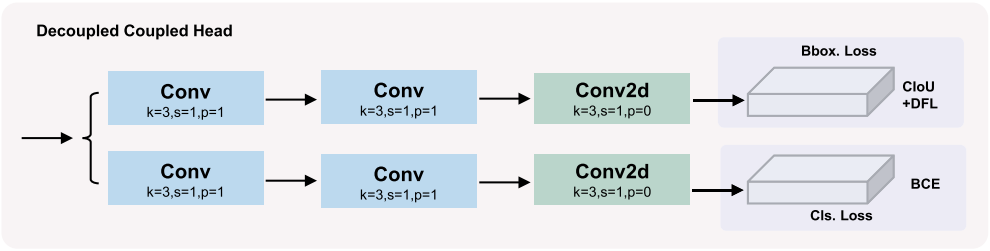


Fig. 9. Decoupled head for detection.

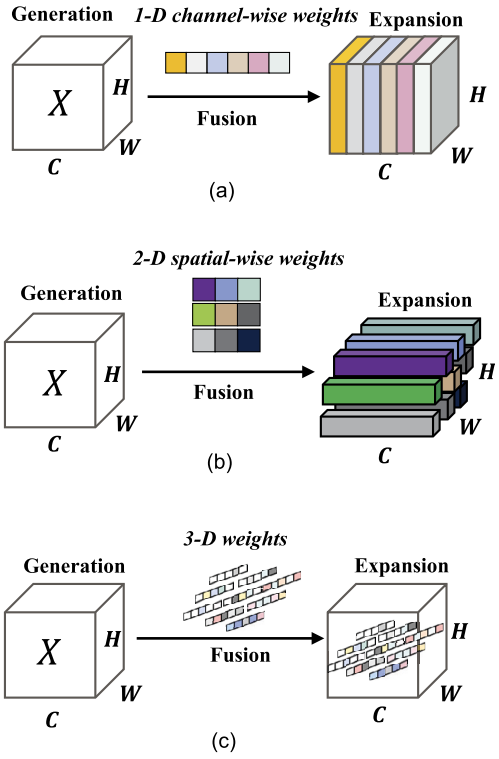


Fig. 10. Workflow of different attention types. Most attention mechanism method generates 1-D or 2-D weights from input feature X then expands the mixed weights for channel (a) and spatial (b). SimAttention directly produces 3-D weights (c).

cient for backbone to learn the discriminative information cues. Attention mechanism could be regarded as a weight selection during feature forward propagation. SimAttention produced 3-D attention weights for the feature maps instead of 1-D or 2-D weights generated by single or spatial attention methods, which treat neurons in each channel or spatial location independently and will limit their capability of learning more distinguishable features. The calculation of SimAttention can be expressed by,

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i + b_t)^2) + (1-(w_t t + b_t))^2 + \lambda w_t^2 \tag{12}$$

$$w_t = -\frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \tag{13}$$

$$b_t = -\frac{1}{2}(t+\mu_t)\, w_t \tag{14}$$

where $\mu_t \sum_{i=1}^{M-1} x_i$ and σ_t^2 are mean and variance feature of that channel.

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{15}$$

where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$ and $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2$. The final process of SimAttention is expressed by:

$$\hat{X} = sigmiod(\frac{1}{\mathbf{E}}) \odot X \tag{16}$$

where $\mathbf{E}$ groups all e_t^* across channel and spatial dimension.

To build Attention Refined Feature Module (ARFM), the SimAttention is inserted before detection heads for refining multi-scale fusing features through generating 3-D weights with well-established channel combined spatial target information. Therefore, THDet detector heads could accumulate the compensatory channel and spatial invariance feature properties to effectively boost the detection performance.

4. Experiments

4.1. Datasets

To verify the effectiveness of our proposed THDet in traffic helmet object detection, we adopt the *HelmetVest Computer Vision Project* [35] from the website of the *robflow*, which was the collection of open source computer vision dataset and APIs. This project includes 4000 images for traffic safety scenario, 3202 images for training and 798 images for validation and test. In order to further distinguish all these images in a more detailed way, we categorized them into four classes: helmet, no helmet, reflective vest and ordinary clothes. The size of the object in this dataset is shown in Fig. 11. There are 2760 medium size objects in this dataset, which accounts a large marge in the all size scope. Besides, the number of the object size of small, large and jumbo are 545, 569 and 126 respectively. And the aspect ration distribution figure indicates that all the objects in the images are balanced for the both large and small traffic safety detection.

4.2. Evaluation standards for detection performance

We employ the following five standards to evaluate and compare the detection performance of our proposed THDet: Precision Rate (P), Recall Rate (R), $mAP_{0.5}$, $mAP_{0.5-0.95}$, Parameters (params.), FLOPs(G) and FPS. All these multiple criteria to measure the performance of the object detection vision tasks.

$$P = \frac{TP}{TP+FP} \tag{17}$$

$$R = \frac{TP}{TP+FN} \tag{18}$$

$$AP_i = \int_0^1 P_i(R_i)\, dR_i \tag{19}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{20}$$

where a threshold is set to determine if the detection is correct. Concretely, the TP, FP and FN denotes of the true positive, false positive,

Table 1
Overall comparison with the mainstream object detection algorithms for traffic helmet object detection.

Methods	*Precision*	*Recall*	$mAP_{0.5}$	$mAP_{0.5-0.95}$	*Params.*	$FLOPs(G)$	FPS
Faster RCNN [16]	0.696	0.753	0.769	0.37	60130340	282.75	78
CascadeRCNN [36]	0.717	0.743	0.769	0.38	68935722	244.11	61
SSD512 [18]	0.796	0.727	0.72	0.386	24146894	34.43	152
RetinaNet [33]	0.582	0.646	0.683	0.308	36166728	205.69	96
FCOS [37]	0.832	0.443	0.675	0.308	31844622	78.69	83
CenterNet [38]	0.782	0.594	0.621	0.252	14431961	19.34	282
RT-DETR [22]	0.791	0.728	0.767	0.395	32003672	108	57
YOLOXs [39]	0.738	0.803	0.822	0.419	8938843	13.32	181
YOLOv5n [40]	0.819	0.776	0.81	0.417	1764577	4.1	303
YOLOv5s [40]	0.831	0.781	0.808	0.426	7020913	15.8	303
YOLOv6n [41]	0.396	0.576	0.814	0.429	4630000	11.34	241
YOLOv6s [41]	0.502	0.581	0.786	0.431	18500000	45.7	215
YOLOv7_tiny [42]	0.797	0.761	0.804	0.388	6015714	13	156
YOLOv7 [42]	0.814	0.803	0.823	0.433	36497954	103.2	94
YOLOv8n [29]	0.792	0.78	0.813	0.433	3006428	8.1	303
YOLOv8s [29]	0.824	0.774	0.814	0.438	11134520	28.2	222
YOLOv9c [43]	0.816	0.758	0.809	0.428	25532300	103.7	125
YOLOv10n [17]	0.766	0.709	0.76	0.387	2695976	8.2	195
THDet	0.815	**0.805**	**0.823**	**0.447**	2002120	7.3	295

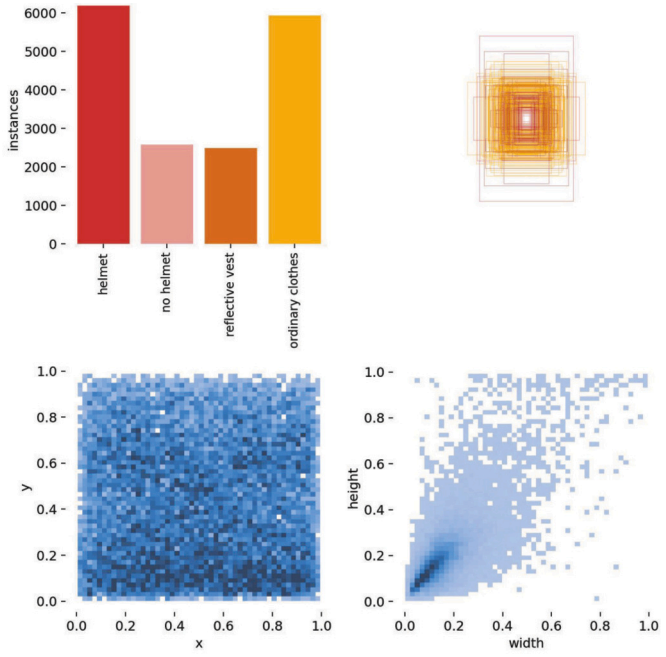


Fig. 11. Visualization of data description.

and false negative bounding box samples, respectively. Average Precision (AP) means the region of the graph is bounded by the P-R curve and the coordinate axis. Mean Average Precision (mAP) indicates the average precision mean of all categories.

4.3. Experimental settings

The experimental configuration used in work is under CPU Intel Core i7-13700KF, GPU Nvidia RTX 4090 24G, Ubuntu 22.04 system, with Pytorch 2.0.1 and Python 3.10.4. The workflow of THDet is shown in the Fig. 12, the input data of our network are RGB images, we scale the image size to 640×640, learning rate 0.01, momentum 0.937, weight decay 5e-4 and 200 epochs for the whole training time with the first 3 epochs for warm-up, which could speeds up the model convergence and help to reduce the model overfitting in the early period of training process.

4.4. Analysis of Experimental Results

Table 1 lists comparison of our proposed THDet with mainstream real-time object detectors. The representative two-stage detectors like Faster RCNN and Cascade RCNN get inferior results compared with our THDet, which realize 0.37 and 0.38 at $mAP_{0.5-0.95}$ respectively, with more than 60 million parameters and 200 GFLOPs model complexity, indicating that this kind type detectors is not suitable for real-time detection. As the typical one-stage detector, SSD algorithm due to the one-input for prediction strategy achieves far excellent detection result 0.386 at $mAP_{0.5-0.95}$ and 152 FPS compared with its counterpart RetinaNet, which the later one exhibits more intricate architecture. We also conduct experiment with anchor-free style detectors like FCOS and CenterNet, both of them perform unpleasant results towards all evaluation metrics except FLOPs and inference speed, explaining that this kind of detectors is inefficient for this task. Our THDet realizes 13.2% improvement at $mAP_{0.5-0.95}$ compared with the transformer-based detection method RT-DETR with more than 30 million parameters, we believe this phenomenon may come from that transformer architecture benefiting from large-scale dataset, so that they require to learn rich features to feed back to their network, our experimental dataset doesn't meet the standard of that scale like COCO [5]. THDet also achieves a leading detection level at $mAP_{0.5-0.95}$ compared with YOLO series algorithms, realizing 7.2%, 4.2%, 15.2% and 3.2% at $mAP_{0.5-0.95}$ improvement compared with YOLOv5n, YOLOv6n, YOLOv7-tiny and YOLOv8n respectively. Meanwhile, THDet also saved 33.4% parameters and reduced 9.9% GFLOPs regarding model size and complexity than YOLOv8n. Compared with YOLOv9c, our proposed method also gets excellent performance not only achieving 4.4% at $mAP_{0.5-0.95}$ increase, but also the parameters and GFLOPs simple accounts for 7.8% and 7.03% of YOLOv9c counterpart. Experimental results of the latest YOLO series detection algorithm YOLOv10n show that THDet also presents leading performance at all the evaluation metrics, indicating that the intrinsic architecture of YOLOv10 is complicated and not efficient for processing dataset used in this helmet detection task. THDet presents high detection performance trade-off with inference speed at FPS 295 with 2 million magnitude parameters alongside 7.3 GFLOPs model complexity, which overwhelmingly outperforms mainstream object detectors, manifesting that THDet is competent to operate fast and accurate helmet detection.

It's known that the activation function plays a very crucial role in neural networks by learning the abstract feature through non-linear transformations. Table 2 shows the results with different choices of activation function in the convolutional blocks. We can apparently see that ConvSP accomplishes the leading detection performance among the rest counterpart on the evaluation metrics with *Precision*, $mAP_{0.5}$

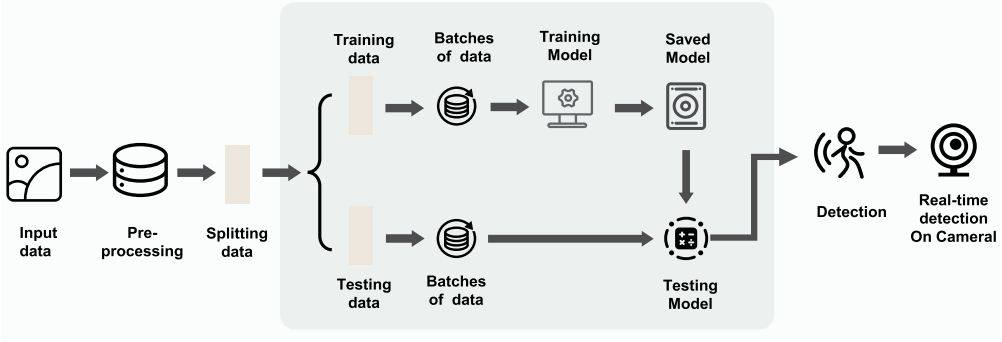


Fig. 12. Workflow of THDet training and testing.

Table 2
Comparison of different activation functions toward CAF convolutional blocks. ConvSL means convolution block with SiLU activation function. ConvRL: with ReLU, ConvGL: with GELU, ConvMS: with Mish, ConvHS: with Hardswish, ConvLR: conv with LeakyReLU, ConvSD: with sigmoid, ConvSP: with Softplus.

Methods	*Precision*	*Recall*	$mAP_{0.5}$	$mAP_{0.5-0.95}$
ConvSL	0.792	0.78	0.813	0.433
ConvRL	0.821	0.758	0.822	0.433
ConvGL	0.815	0.763	0.82	0.43
ConvMS	0.808	0.779	0.82	0.433
ConvHS	0.822	0.77	0.818	0.428
ConvLR	0.8	0.768	0.814	0.428
ConvSD	0.795	0.75	0.806	0.414
ConvSP	**0.832**	0.768	**0.824**	**0.434**

Table 3
Enhanced C2f module with attention for feature extraction.

Methods	*Precision*	*Recall*	$mAP_{0.5}$	$mAP_{0.5-0.95}$	*Params.*	$FLOPs(G)$	*FPS*
SE [44]	0.81	0.77	0.817	0.431	3012988	8.1	208
ECA [45]	0.809	0.787	0.819	0.43	3006485	8.1	278
CBAM [46]	0.795	0.787	0.814	0.433	3060512	8.1	256
ESE [47]	0.806	0.769	0.814	0.428	3059532	8.1	250
CA [30]	0.815	0.77	0.819	**0.436**	3022892	8.1	213

and $mAP_{0.5-0.95}$ with 0.832, 0.824 and 0.434 respectively. The smooth nature of softplus activation function facilitates the differentiability of feature matrix computation, just shown in Fig. 2 (the last one in the second row). On the other hand, we speculate the reason of this experiment result may lie in that softplus activation function is much suitable for dataset in this task due the simplicity and distinguishable features of helmet and reflective vest. Furthermore, sigmoid activation function shows the worst detection performance, explaining that it's not suitable for helmet detection.

Table 3 shows the feature extraction contrast experiments results when attention mechanisms employed in the C2f blocks. It's clear that CA achieves the best score 0.436 at $mAP_{0.5-0.95}$ with 213 FPS. Compared with channel attention style mechanism SE and ECA, CA realize the 1.2% and 1.4% improvement at $mAP_{0.5-0.95}$ respectively. CA enhances the feature extraction ability by channel and spatial dimension, it adaptively selects both important objects and regions and leverages the global spatial information for following neural networks with enhanced discrimination of features. Combined with another channel and spatial branch attention, the experimental results of CBAM are slightly inferior than CA, but perform excellent on FPS with 256. We finally choose the CA to enhance the extraction ability of C2f block considering the better trade-off of its detection accuracy and inference speed.

Table 4 lists the experimental results of different IoU variants loss function for bounding box regression. It's notable that CIoU achieves the leading overall detection accuracy 0.433 at $mAP_{05-0.95}$ with 303 FPS among other methods, indicating that CIoU is more efficient when dealing with overlap area of predicted proposals and ground truth. We have also tested the Focal Loss strategy with all the IoU-based methods. Detection improvement can be gained with Focal method only for GIoU and CIoU, which is 0.434 and 0.436 at $mAP_{0.5-0.95}$ with the same 313 FPS, implying Focal loss exhibits positive role in dealing with imbalanced samples during training. Considering the excellent results of CIoU, our THDet adopts the Focal_CIoU as bounding box regression loss function.

In terms of the influence of the detection heads to the whole model performance, we conduct experiments to validate our hypothesis whether with more detection heads will boost the final results. The impact of the heads numbers from different stages is shown in Table 5, we can see that when using 4 heads for detection, the performance of $mAP_{0.5-0.95}$ has dropped (like 0.426 and 0.428) compared with baseline model 0.433 at $mAP_{0.5-0.95}$. Besides, the inference speed and model complexity also deteriorate, which indicates that when adding an extra small detection head or large object detection head have negative effects to the final results and the model became extreme redundancy for recognizing the relative distinguishable helmet features. By employing 2 detection heads with P3 & P4 accomplish the best balance regarding to the evaluation of parameters, model complexity and inference speed, which also saved the 33.7% parameters and 13.9% FPS boost with 345 compared with baseline. The experiments testify that 2 proper combined detection heads are the key to achieve the lightweight model design and reach real-time detection in the filed of traffic helmet detection scenario.

Table 4
Experimental results of IoU variants loss function for bounding box regression.

Methods	*Precision*	*Recall*	$mAP_{0.5}$	$mAP_{0.5-0.95}$	*FPS*
CIoU [31]	0.792	0.77	0.814	0.433	303
DIoU [48]	0.815	0.782	0.823	0.43	294
SIoU [49]	0.816	0.761	0.813	0.428	303
EIoU [50]	0.808	0.772	0.815	0.428	303
GIoU [32]	0.816	0.769	0.822	0.432	303
Focal_DIoU	0.805	0.787	0.818	0.428	313
Focal_EIoU	0.809	0.773	0.821	0.428	313
Focal_SIoU	0.809	0.773	0.818	0.43	313
Focal_GIoU	0.815	0.776	0.818	0.434	313
Focal_CIoU	**0.82**	0.778	0.82	**0.436**	313

Table 5
Experimental results about the choices of detection heads. SOB: small object detection head. LOB: large object detection head. Hs: the number of the detection heads.

Methods	*Precision*	*Recall*	$mAP_{0.5}$	$mAP_{0.5-0.95}$	*Params.*	*FLOPs*(*G*)	*FPS*
BaseLine 3 Hs	0.792	0.78	0.813	0.433	3006428	8.1	303
+ SOB 4Hs P2	0.829	0.751	0.816	0.426	2921568	12.1	244
+ LOB 4Hs P6	0.795	0.764	0.822	0.428	4864608	8.2	-
2 Hs P3 & P5	0.791	0.789	0.807	0.425	2780568	7.4	303
2 Hs P3 & P4	0.789	**0.791**	0.818	0.428	**1993240**	**7.3**	**345**

Table 6
Optimization for detection headers for more effective refined features.

Type	Method	*Precision*	*Recall*	$mAP_{0.5}$	$mAP_{0.5-0.95}$	*Params.*	*FLOPs*(*G*)	*FPS*
3 Heads	RFAConv [51]	0.829	0.751	0.816	0.426	2921568	12.1	244
	EVC [52]	0.803	0.78	0.815	0.431	8792092	18.3	138
	SEv2 [53]	0.823	0.768	0.815	0.426	3032028	8.1	182
2 Heads	RepVGG [54]	0.791	0.789	0.807	0.425	2780568	7.4	303
	EVC [52]	0.799	0.779	0.814	0.428	3343768	14	172
	SimAttention [34]	0.814	0.773	0.815	**0.433**	**1994320**	**7.3**	**280**

We also have conducted the experiments to compare the detection performance regarding the numbers of heads when constructing the ARFM. For the three detection heads style, we employ three different convolutional attention method to build cost-efficient detection head. According to Table 6, it's obvious that EVC block achieves the best detection accuracy 0.431 at $mAP_{0.5-0.95}$ by prioritizing spatial features to balance convolution kernel parameter sharing. However, EVC block has more than 8 million parameters and performs at low inference speed 138 FPS, becoming a burden for economic lightweight model design. For utilizing 2 heads for detection, we exploit three different approaches to build ARFM. We can find that the RepVGG method realizes the fastest running speed at 303 FPS, but detection accuracy is 1.9% lower at $mAP_{0.5-0.95}$ that SimAttention. EVC block performs a little pleasant detection result compared with RepVGG but still 1.2% lower than SimAttention at $mAP_{0.5-0.95}$. Among all efficient-building methods, SimAttention reaches the best detection results 0.433 at $mAP_{0.5-0.95}$ and 280 FPS inference speed with 1994320 parameters and only 7.4 GFLOPs model complexity, revealing that SimAttention is parameter-efficient and an easy plug-and-play method to build high performance detection head.

To verify the effectiveness of our proposed method, we have also conducted the ablation study to verify the contributions of each designed modules. Table 7 shows that after using the 2 Heads for detection, the parameters drop by 33.7% with very high inference speed at 345 FPS compared with baseline model. When combining the Softmax activation function and CA mechanism into the C2f blocks, 1.4% improvements are achieved at $mAP_{0.5-0.95}$ with little parameters increase, indicating this integrated convolutional blocks design is effective to extract helmet features during training. ARFM improved the performance by 0.7% with solely 2 heads for detection, which mitigates the accuracy loss at $mAP_{0.5-0.95}$ to some extent. When combined all the metrics together, our proposed model reach the accuracy 0.447 at $mAP_{0.5-0.95}$, realizing 3.2% overall detection improvements in comparison with baseline model. Meanwhile, our method also achieves the lightweight network scheme with only 2002120 parameters and 7.3 GFLOPs model complexity as well as 295 FPS inference speed, realizing the standard of end-to-end algorithm philosophy and real-time detection.

To further validate robustness of THDet on different running conditions, we conduct experiments to test THDet with prevailing real-time detectors on GPU and CPU separately. The experiments are performed under GPU RTX 4090 and CPU i7-13700KF and the results are shown in Table 8. It's obvious that THDet produces the remarkable detection performance both on GPU and CPU with FPS 295 and 51 respectively. Although YOLOv5n reaches high inference detection speed on GPU 404 FPS and CPU 68 FPS due to its simple and efficient architecture, there still is 7.2% accuracy lagging behind at $mAP_{0.5-0.95}$ compared with THDet. RT-DETR owes the running speeding at only 5 FPS on the CPU platform, explaining that transformer-based method needs high advanced computation resources like sophisticated GPU hardware. Additionally, THDet also benefits from lightweight network design with only 2002120 parameters and 7.3 GFLOPs, in comparison with heavy models like YOLOv8s and YOLOv9c, which have more than 5, 12 times parameters than THDet, implying that parameter-burden models cannot meet the standard of real-time object detector. Compared with the latest released YOLO series detector YOLOv10n, THDet achieves 15.5% notable leading improvement at $mAP_{0.5-0.95}$. Meanwhile, THDet outperforms YOLOv10n regarding to the model parameters and model complexity as well. Especially, THDet operates with much higher inference speed both on GPU and CPU devices than YOLOv10n.

Table 7
Ablation study of the improvement modules.

Baseline	LightHead	Softmax	Focal_CIoU	CA_C2f	ARFM	$mAP_{0.5-0.95}$	*Params.*	$FLOPs(G)$	*FPS*
✓						0.433	3006428	8.1	303
✓	✓					0.428	1993240	7.3	345
✓		✓				0.434	3006428	8.1	303
✓			✓			0.436	3006428	8.1	313
✓		✓		✓		0.439	3022892	8.1	285
✓	✓				✓	0.431	1993240	7.3	295
✓	✓	✓	✓	✓	✓	0.447	2002120	7.3	295

Table 8
Performance of different algorithms towards complexity and inference on the GPU and CPU Platform respectively.

Methods	$mAP_{0.5-0.95}$	*Params.*	$FLOPs(G)$	*FPS_GPU*	*FPS_CPU*
RT-DETR [22]	0.395	32993672	108	57	5
YOLOv5n [40]	0.417	1764577	4.1	404	68
YOLOv5s [40]	0.426	7020913	15.8	303	25
YOLOv6n [41]	0.429	4630000	11.34	241	-
YOLOv7t [42]	0.394	6015714	13	333	11
YOLOv8n [29]	0.433	3006428	8.1	303	47
YOLOv8s [29]	0.438	11134520	28.2	222	20
YOLOv9c [43]	0.428	25532300	103.7	125	5
YOLOv10n [17]	0.387	2695976	8.2	195	33
THDet	0.447	2002120	7.3	295	51

Fig. 13. Visualization of comparison of YOLOv8n and THDet for traffic helmet detection.

4.5. Analysis of Detection Results

Several images from the dataset were chosen to verify the effectiveness of model in the real application scenarios. Fig. 13 and Fig. 14 shows the detection results of THDet and baseline YOLOv8n algorithm in different human number conditions. It can be found from the pictures that the detection confidence of THDet is generally higher than the YOLOv8n counterpart, which proves that THDet exhibits excellent detection performance by employing the proposed improvement metrics. In the meanwhile, THDet also have advantages detecting helmet and reflective vest in the crowded environment with precise bounding box localization and helmet classification. To summarize, THDet outperforms YOLOv8n when detecting smaller-scale with various shaped-helmet objects and having stronger generalization, leading to more precise helmet detection and recognition.

5. Conclusion

In this work, we design an efficient and real-time object detector for traffic helmet detection named THDet. Several strategies have been adopted to improve the detection performance. First, we employ Coordinate Attention and Softplus activation function embedded into C2f blocks to construct powerful feature extraction backbone network. Next, CIoU combined with Focal Loss method together named Focal_CIoU as the bounding box loss function for reducing missed detection of overlap targets and balancing the positive and negative training examples. Then, two detection heads pattern is introduced and validated the feasibility to realize lightweight acceptable detection results with vast model parameters and complexity drop. Finally, we build Attention Refined Features Module (ARFM) with SimAttention with 3-D weights both from spatial and channel dimension to further refine and calibrate the multi-scales fused features for better distinguishable expression. The detection eval-

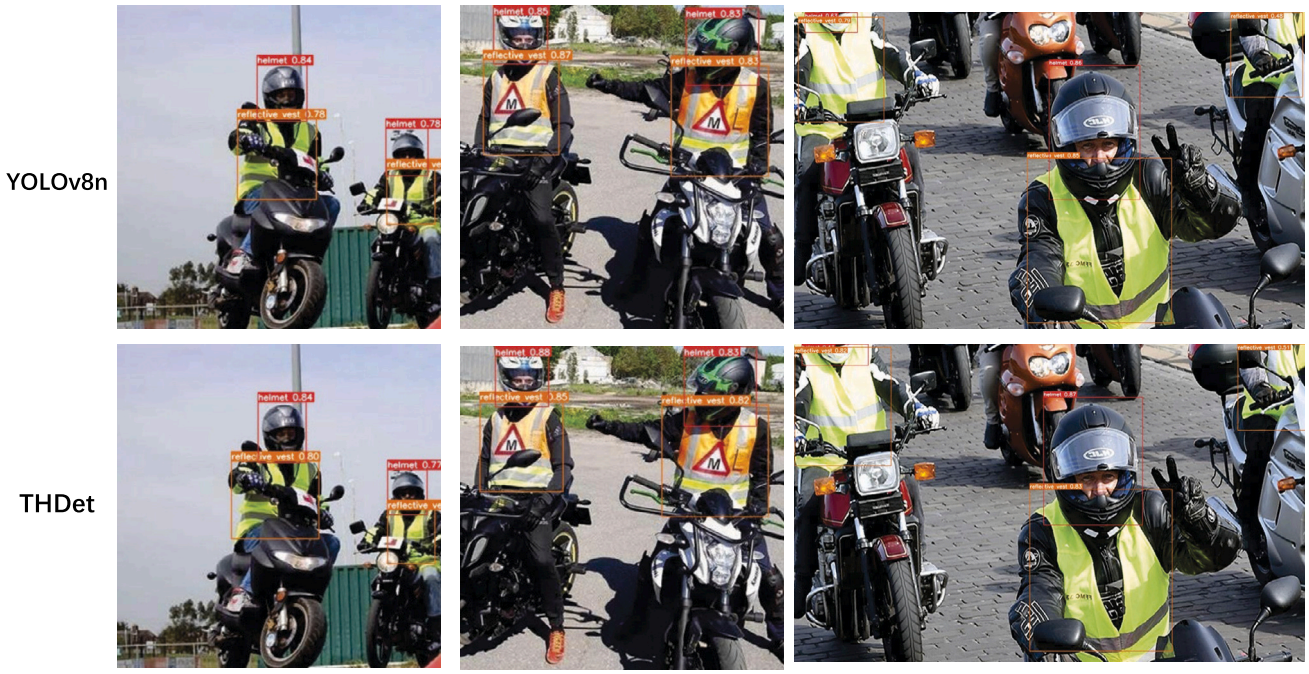


Fig. 14. More examples of detection visualization of YOLOv8n and THDet for traffic helmet detection.

uation metric on $mAP_{0.5-0.95}$ of THDet is 0.447 and FPS is 295, a 3.2% detection accuracy improvement and 33.4% parameters drop compared baseline YOLOv8n, showcasing THDet reaches excellent detection performance and has the potential for real-time detection and application. In the future, we will further optimize the training process, inference time and model deployment of THDet on the edge devices with limited computation resource.

Abbreviations

The following abbreviations are used in this manuscript:

DCNN	Deep Convolutional Neural Networks
YOLO	You Look Only Once
RPN	Region Proposal Networks
FPN	Feature Pyramid Networks
C2f	CSPDarknet53 to 2-Stage FPN
C3	CSP Bottleneck with 3 Convolutions
CA	Coordinate Attention
CA_C2f	Coordinate Attention with C2f
IoU	Intersection over Union Loss Function
CIoU	Complete-IoU
Focal_CIoU	Focal Loss Function with CIoU
ARFM	Attention Refined Features Module
LightHead	Lightweight detection Head
NMS	Non-Maximum Suppression
BBR	Bounding Box Regression
P	Precision
R	Recall
Params.	Parameters
FLOPs	Floating Point Operations
FPS	Frames Per Second

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376252, 61976196, U22A20102); Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003).

References

[1] Z. Zou, K. Chen, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: a survey, Proc. IEEE 111 (3) (2023) 257–276.
[2] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[3] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[7] C.-Y. Wang, H.-Y.M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, I.-H. Yeh, Cspnet: a new backbone that can enhance learning capability of cnn, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 390–391.
[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[9] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
[10] J.-S. Lim, M. Astrid, H.-J. Yoon, S.-I. Lee, Small object detection using context and attention, in: 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE, 2021, pp. 181–186.
[11] C. Feng, Y. Zhong, Y. Gao, M.R. Scott, W. Huang, Tood: task-aligned one-stage object detection, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, 2021, pp. 3490–3499.

[12] Z. Ge, S. Liu, Z. Li, O. Yoshie, J. Sun, Ota: Optimal transport assignment for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 303–312.
[13] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, S.-H.G. Chan, Run, don't walk: chasing higher flops for faster neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12021–12031.
[14] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: more features from cheap operations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1580–1589.
[15] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2010) 303–338.
[16] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[17] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, Yolov10: real-time end-to-end object detection, arXiv preprint, arXiv:2405.14458, 2024.
[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.
[19] S. Liu, D. Huang, et al., Receptive field block net for accurate and fast object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 385–400.
[20] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, H. Ling, M2det: a single-shot object detector based on multi-level feature pyramid network, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9259–9266.
[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[22] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, Y. Liu, Detrs beat yolos on real-time object detection, arXiv preprint, arXiv:2304.08069, 2023.
[23] J. Wang, Y. Chen, Z. Dong, M. Gao, Improved yolov5 network for real-time multi-scale traffic sign detection, Neural Comput. Appl. 35 (10) (2023) 7853–7865.
[24] J. Lian, Y. Yin, L. Li, Z. Wang, Y. Zhou, Small object detection in traffic scenes based on attention feature fusion, Sensors 21 (9) (2021) 3031.
[25] G. Zeng, W. Huang, Y. Wang, X. Wang, E. Wenjuan, Transformer fusion and residual learning group classifier loss for long-tailed traffic sign detection, IEEE Sens. J. (2024).
[26] J.-Y. Lee, W.-S. Choi, S.-H. Choi, Verification and performance comparison of cnn-based algorithms for two-step helmet-wearing detection, Expert Syst. Appl. 225 (2023) 120096.
[27] H. Li, D. Wu, W. Zhang, C. Xiao, Yolo-pl: Helmet wearing detection algorithm based on improved yolov4, Digit. Signal Process. (2023) 104283.
[28] J. Mi, J. Luo, H. Zhao, X. Huang, Improved dense residual network with the coordinate and pixel attention mechanisms for helmet detection, Int. J. Mach. Learn. Cybern. (2024) 1–17.
[29] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLO https://github.com/ultralytics/ultralytics, jan 2023.
[30] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.
[31] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, W. Zuo, Enhancing geometric factors in model learning and inference for object detection and instance segmentation, IEEE Trans. Cybern. 52 (8) (2021) 8574–8586.
[32] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: a metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 658–666.
[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[34] L. Yang, R.-Y. Zhang, L. Li, X. Xie, Simam: a simple, parameter-free attention module for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2021, pp. 11863–11874.
[35] Data, Helmetvest dataset, https://universe.roboflow.com/data-u4eek/helmetvest, visited on 2024-04-11 (oct 2022), https://universe.roboflow.com/data-u4eek/helmetvest.
[36] Z. Cai, N. Vasconcelos, Cascade r-cnn: delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.
[37] Z. Tian, C. Shen, H. Chen, T. He, Fcos: fully convolutional one-stage object detection, arxiv 2019, arXiv preprint, arXiv:1904.01355, 1904.
[38] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: keypoint triplets for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6569–6578.
[39] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: exceeding yolo series in 2021, arXiv preprint, arXiv:2107.08430, 2021.
[40] G. Jocher, Yolov5, by ultralytics (may 2020), https://doi.org/10.5281/zenodo.3908559, https://github.com/ultralytics/yolov5.
[41] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., A single-stage object detection framework for industrial applications, Yolov6, arXiv preprint, arXiv:2209.02976, 2022.
[42] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.
[43] C.-Y. Wang, I.-H. Yeh, H.-Y.M. Liao, Yolov9: learning what you want to learn using programmable gradient information, arXiv preprint, arXiv:2402.13616, 2024.
[44] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[45] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.
[46] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[47] Y. Lee, J. Park, Centermask: real-time anchor-free instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13906–13915.
[48] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-iou loss: faster and better learning for bounding box regression, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 12993–13000.
[49] Z. Gevorgyan, Siou loss: more powerful learning for bounding box regression, arXiv preprint, arXiv:2205.12740, 2022.
[50] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, T. Tan, Focal and efficient iou loss for accurate bounding box regression, Neurocomputing 506 (2022) 146–157.
[51] X. Zhang, C. Liu, D. Yang, T. Song, Y. Ye, K. Li, Y. Song, Rfaconv: innovating spatial attention and standard convolutional operation, arXiv preprint, arXiv:2304.03198, 2023.
[52] Y. Quan, D. Zhang, L. Zhang, J. Tang, Centralized feature pyramid for object detection, IEEE Trans. Image Process. (2023).
[53] M. Narayanan, Senetv2: aggregated dense layer for channelwise and global representations, arXiv preprint, arXiv:2311.10807, 2023.
[54] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, Repvgg: making vgg-style convnets great again, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13733–13742.

Yi Li received the M.S. degree in Software Engineering from College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua China, in 2021, where he is currently pursuing the Ph.D. degree in Computer Science and Technology. He is a student member of CFF. His current research interests include deep learning, computer vision tasks, Object Detection, Instance Segmentation, Object Tracking.

Huiying Xu received the M.S. degree from National University of Defense Technology (NUDT), China. She is an associate professor with the School of Computer Science and Technology, Zhejiang Normal University, and also the researcher of Research Institute of Ningbo Cixing Co. Ltd, China. Her research interests include Kernel learning and feature selection, Object Detection, Vision SLAM, Computer vision, Image processing, Pattern recognition, Computer simulation, Deep clustering, Generative Adversarial Network, Diffusion Model, Clustering Ensemble, Multiple Kernel Learning, Learning with incomplete data and their applications. She is a member of the China Computer Federation. She has published papers, including those in highly regarded journals such International Journal of Intelligent Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Multimedia, etc.

Xinzhong Zhu received the Ph.D. degree from Xidian University and M.S. degree from National University of Defense Technology (NUDT), China. He is a professor with the School of Computer Science and Technology, Zhejiang Normal University, and also the chief scientist of Beijing Geekplus Technology Co., Ltd. and president of Research Institute of Ningbo Cixing Co., Ltd., China. His research interests include Machine learning, Deep clustering, Computer vision, Object detection, Segmentation, Recognition and Tracking, Diffusion Model, Manufacturing informatization, Manufacturing informatization, Robotics and System integration, Laser SLAM, Vision SLAM, Low Quality Data Learning, Multiple Kernel Learning, and Intelligent manufacturing. He is a member of the ACM and certified as CCF distinguished member. Dr. Zhu has published more than 30 peer-reviewed papers, including those in highly regarded journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Multimedia, the IEEE Transactions on Knowledge and Data Engineering, CVPR, NeurIPS, AAAI, IJCAI, etc. He served on the Technical Program Committees of IJCAI 2020 and AAAI 2020.

Xiao Huang received a Ph.D. degree from East China Normal University. She is the Dean of the College of Education, the Joint Education institute of Zhejiang Normal University and University of Kansas, She has worked as a professor in 2016, and also served as a PhD Supervisor and the Director of Science Education Research Center. She is the chief expert of Research Institute of Education Reform and Development in Zhejiang Philosophy and Social Sciences Key Cultivation Research Base, expert of international ISO standard TC/286/WG 4 for school-enterprise cooperation. Her research fields include STEM education, Nature of science and Scientifie inquiry, She is the member of NARST(National Association for Research in Science Teaching), ESERA(European Science Education Research Association) and AAPT(American Association of Physics Teachers).

Hongbo Li, received his Ph.D. degree in computer science from Tsinghua University in 2009. Currently, he holds the position of the Chief Technology Officer and Co-founder of Beijing Geek+ Technology Co., Ltd China. In addition, he also serves as the secretary-general of Chinese Intelligent service Society and is an Editorial Board Member of several high-profile journals. His research interests include the design and application of intelligent robots, intelligent information process, and intelligent logistic systems. He has published more than 70 papers in prestigious journals and conference, and has been awarded more than 120 patents, including 46 international invention patents.