Neural Processing Letters (2024) 56:90
https://doi.org/10.1007/s11063-024-11558-4

# HF-YOLO: Advanced Pedestrian Detection Model with Feature Fusion and Imbalance Resolution

**Lihu Pan[1] · Jianzhong Diao[1] · Zhengkui Wang[2] · Shouxin Peng[1] · Cunhui Zhao[3]**

Accepted: 11 February 2024


**Abstract**
Pedestrian detection is crucial for various applications, including intelligent transportation and video surveillance systems. Although recent research has advanced pedestrian detection models like the YOLO series, they still face limitations in handling diverse pedestrian scales, leading to performance challenges. To address these issues, we propose HF-YOLO, an advanced pedestrian detection model. HF-YOLO tackles the complexities of pedestrian detection in complex scenes by addressing scale variations and occlusions among pedestrians. In the feature fusion stage, our algorithm leverages both shallow localization information and deep semantic information. This involves fusing P2 layer features and adding a high-resolution detection layer, significantly improving the detection of small-scale pedestrians and occluded instances. To enhance feature representation, HF-YOLO incorporates the HardSwish activation function, introducing more non-linear factors and strengthening the model's ability to represent complex and discriminative features. Additionally, to address regression imbalance, a balance factor is introduced to the CIoU loss function. This modification effectively resolves the imbalance problem and enhances pedestrian localization accuracy. Experimental results demonstrate the effectiveness of our proposed algorithm. HF-YOLO achieves notable improvements, including a 3.52% increase in average precision,

Jianzhong Diao, Zhengkui Wang, Shouxin Peng and Cunhui Zhao have contributed equally to this work.

✉ Lihu Pan
panlh@tyust.edu.cn

Jianzhong Diao
diaojianzhong@stu.tyust.edu.cn

Zhengkui Wang
zhengkui.wang@singaporetech.edu.sg

Shouxin Peng
psx19980801@outlook.com

Cunhui Zhao
13903402412@163.com

1 School of Computer Science and Technology, Taiyuan University of Science and Technology, 63 Waliu Rd, Taiyuan 030024, Shanxi, China

2 ICT Cluster, Singapore Institute of Technology, 10 Dover Drive, Singapore 139651, Singapore

3 Jingying Shuzhi Technology Co.,Ltd., 103 Changzhi Rd, Taiyuan 030012, Shanxi, China

Published online: 06 March 2024

a 1.35% boost in accuracy, and a 4.83% enhancement in recall. Moreover, the algorithm maintains real-time performance with a detection time of 8.5ms, meeting the stringent requirements of real-time applications.

**Keywords** Pedestrian detection · Object detection · Activation function · YOLO · Loss function

## 1 Introduction

Pedestrian detection is a widely studied object detection problem that finds extensive applications in domains such as intelligent video surveillance [1], intelligent transportation [2], and autonomous driving systems [3, 4]. It also serves as a fundamental technology supporting tasks like pedestrian pose estimation [5, 6] and pedestrian re-identification [7, 8]. The accuracy of pedestrian detection algorithms directly impacts the performance of these tasks.

In real-world, the diversity in pedestrian poses, varying scales, and environmental factors, including occlusions, present significant challenges to detection algorithms, necessitating robust and precise solutions.

Traditional pedestrian detection methods, such as Viola–Jones [9], histogram of oriented gradients (HOG) [10], and scale-invariant feature transform (SIFT) [11], rely on manually designed features and template matching techniques. While these methods are straightforward to implement, their performance and generalization capacity are limited by the constraints of handcrafted feature engineering, often resulting in suboptimal outcomes when faced with complex scenarios.

In recent years, considerable progress has been made in object detection techniques. State-of-the-art algorithms, such as multi-anchor faster R-CNN [12], multiscale attention fusion [13], and multimodal detectors [14], leverage convolutional neural networks (CNNs) to enhance pedestrian detection performance by improving accuracy. Among these approaches, the YOLO (you only look once) series has consistently maintained a prominent position, striking a favorable trade-off between detection precision and processing speed.

However, it is noteworthy that YOLO models exhibit certain limitations that hinder their performance in pedestrian detection. Specifically, these models face challenges in effectively handling the diverse scales of pedestrians and addressing the imbalance between positive and negative samples. Primarily, the inherent flexibility and variability in human poses give rise to pedestrians appearing at various scales within images (Fig. 1a). Consequently, accurately detecting pedestrians across these varying scales becomes a complex task. Furthermore, scenarios characterized by dense pedestrian crowds introduce additional difficulties due to occlusions occurring between individuals (Fig. 1b). Such occlusions obstruct the complete visibility of pedestrians, further complicating their accurate detection by YOLO models. In addition to scale and occlusion challenges, the imbalance between positive and negative samples during the target regression process adversely impacts the precision of object localization by YOLO models. This imbalance skews the learning process towards the dominant class, resulting in suboptimal performance when localizing pedestrians. Hence, the task of effectively addressing these limitations and devising resilient solutions continues to be a pivotal and captivating area of research. Resolving these challenges holds paramount importance for advancing the field of pedestrian detection and enabling the development of more accurate and reliable systems.

(a) (b)

**Fig. 1** **a** and **b** showcase the scenarios of pedestrians in real-world environments. In Figure (**a**), the red and green boxes depict pedestrians of varying scales. In Figure (**b**), the blue and orange boxes illustrate instances of occlusion. (Color figure online)

To overcome these challenges, this paper presents a groundbreaking model called HF-YOLO, which addresses the limitations in pedestrian detection performance. Our model introduces innovative techniques to enhance the accuracy of detecting pedestrians with small scales and occlusions. It achieves this by leveraging feature fusion across multiple hierarchical levels, enabling the integration of high-resolution features to capture both high-level semantic information and low-level localization cues. Additionally, a dedicated small object detection layer is introduced to improve the accuracy of detecting pedestrians with varying scales and occlusions.

The contributions of our work can be summarized as follows:

- We propose HF-YOLO, a novel pedestrian detection model that incorporate the HardSwish activation function within the convolutional blocks of our model, enhancing its feature representation capability and introducing greater non-linearity.
- We employ feature fusion across multiple hierarchical levels. This novel approach effectively tackles the challenges associated with small-scale pedestrian detection and occlusions.
- To address the issue of imbalanced high- and low-quality samples in the regression loss function, we introduce a balancing factor. This factor redirects the model's focus towards accurate regression for high-quality samples, overcoming the bias introduced by the larger number of low-quality samples.
- Extensive evaluations are conducted, comparing our proposed model against six baselines. The results demonstrate the superior performance of our model, surpassing existing approaches in terms of detection accuracy and reliability.

The paper's structure is as follows: In Sect. 2, we review prior research on pedestrian detection. Section 3 provides a detailed explanation of the HF-YOLO model. In Sect. 4, we present the dataset used and discuss the experimental results. Finally, in Sect. 5, we conclude the paper.

## 2 Related Work

Currently, object detection algorithms can be broadly classified into two major categories: Anchor-based algorithms and Anchor-free algorithms. Anchor-based algorithms encompass both two-stage approaches (e.g., RCNN [15], Fast R-CNN [16], Faster R-CNN [17]) and one-stage approaches (e.g., SSD [18] and YOLO series [19–21]). In Anchor-based algorithms, a set of predefined anchor boxes are generated on the image at different scales and aspect ratios. These anchor boxes serve as reference regions for detecting objects. Convolutional neural networks are employed to extract features from the image, followed by classification and regression operations performed on each anchor box. Finally, a non-maximum suppression algorithm is applied to filter out redundant detection boxes and obtain the final detection results.

In contrast, Anchor-free methods, such as CenterNet [22] and FCOS [23], adopt a different paradigm. These approaches treat objects as single points during model construction. After feature extraction, specific points (e.g., center points) are selected as key representations for object detection. The classification and regression tasks are decoupled into separate branches, allowing for independent prediction of object presence and spatial localization. Subsequently, a non-maximum suppression technique is applied to eliminate overlapping detections and produce the final detection results. While Anchor-free methods alleviate the need for anchor box generation, they often require a larger number of candidate points during detection. Consequently, these methods necessitate more extensive training data and sophisticated training strategies to achieve optimal detection performance in various scenarios.

The development of object detection has progressed from traditional methods to the era of deep learning-based detection. Early object detection algorithms relied on handcrafted image feature descriptors tailored for specific tasks, aiding in discerning target positions or categories. However, these manually crafted features lacked effective image representations, necessitating the design of more intricate feature representations to obtain high-quality image features. As detection techniques advanced, methods rooted in deep learning emerged, leveraging convolutional neural networks (CNNs) and akin methodologies to acquire robust and sophisticated feature representations. Yu et al. extracted multiple features from input images and an image database. They constructed multiple hypergraph Laplacian operators and formulated sparse codes [24]. Simultaneously, they preserved the locality of the obtained sparse codes by employing manifold learning on the hypergraph. Cao et al. introduced a discriminative region of interest (RoI) pooling scheme that samples from various sub-regions of a proposal and performs adaptive weighting to obtain discriminative features [25]. Yu et al. designed a novel fine-grained image recognition framework [26], introducing a feature selection module to address noise and high dimensionality in features. They incorporated weight vectors with sparse constraints and an improved "RELU" operator. The feature selection model achieved higher accuracy and a larger compression ratio. Woo et al. optimized the network from both channel and spatial perspectives, further enhancing the model's feature extraction effectiveness in both dimensions [27]. Lv et al. proposed the RTDETR model, which utilizes Transformers to process the features extracted from the last layer of the backbone, enhancing the model's ability to differentiate features among various objects [28].

While significant advancements have been made in the field of general object detection algorithms, their direct application to pedestrian detection has proven to be unsatisfactory [29]. Zhang et al. proposed a novel occlusion-aware R-CNN detection algorithm [30], building upon the faster R-CNN framework. This approach incorporates an occlusion processing unit, which effectively captures five distinct partial features of pedestrians. These part-level

features are subsequently combined with the global features of the target, employing a weighted summation technique to obtain a comprehensive pedestrian detection outcome. Liu et al. introduced a density prediction module and an adaptive non-maximum suppression (NMS) method in order to address the challenges associated with pedestrian detection [31]. The adaptive NMS method dynamically adjusts the NMS threshold based on the density of the objects being detected. In scenarios with dense pedestrian presence, the NMS threshold is heightened to ensure a high recall rate. Conversely, in cases with sparser object distributions, the NMS threshold is lowered to alleviate the issue of redundant detection boxes. This adaptive approach effectively resolves the predicament of losing highly overlapped targets or generating false positives due to a fixed threshold. Chu et al. proposed a multi-instance prediction method to handle severe overlap among multiple targets [32]. Instead of predicting a single instance, their method predicts a set of highly overlapped instances for each proposal box. This approach reduces the false negative rate by designing a loss function that minimizes the distance between predicted boxes and ground truth boxes, supervising the learning of instance set prediction.

Xia et al. argued that incorporating multi-scale information can improve the robustness of the network without losing information [33]. They introduced multi-scale dilation residual modules into the backbone network to increase the receptive field and capture more global and higher-level semantic features. Li et al. utilized an improved Res2Net as the backbone network to enhance the model's multi-scale representation capability for pedestrians [34]. Addressing the limited ability of a single feature extraction block to extract semantic information at different levels, Wang et al. proposed Three ResNet Blocks [35]. This module integrates three different basic blocks, each of which extracts pedestrian information, to enhance the information flow in the network structure and improve the accuracy of detection results.

These related works demonstrate the efforts made to overcome the limitations and challenges in pedestrian detection. The proposed methods introduce novel techniques, such as occlusion processing, adaptive NMS, and multi-instance prediction, to enhance the performance of pedestrian detection models. Additionally, advancements in backbone networks, such as Res2Net [36], further contribute to the improvement of multi-scale representation and feature extraction capabilities.

## 3 HF-YOLO Model for Pedestrian Detection

### 3.1 Model Overview

The YOLO series models are a prominent category of one-stage object detection methods. These models combine the tasks of object classification and localization regression by utilizing anchor boxes. This integration allows YOLO models to achieve high efficiency, flexibility, and good generalization performance, making them highly popular in both academia and industry.Wang et al. proposed a real-time detection model called YOLOv7 [37]. It is an advanced version of the YOLO series of real-time object detection models. Building on the success of its predecessor, further optimization has been introduced to improve its detection performance.

To address the challenges associated with pedestrian detection in complex environments, we propose the HF-YOLO model, which builds upon the foundation of YOLOv7. The HF-YOLO architecture, illustrated in Fig. 2, comprises three essential components: Backbone, Neck, and Head.

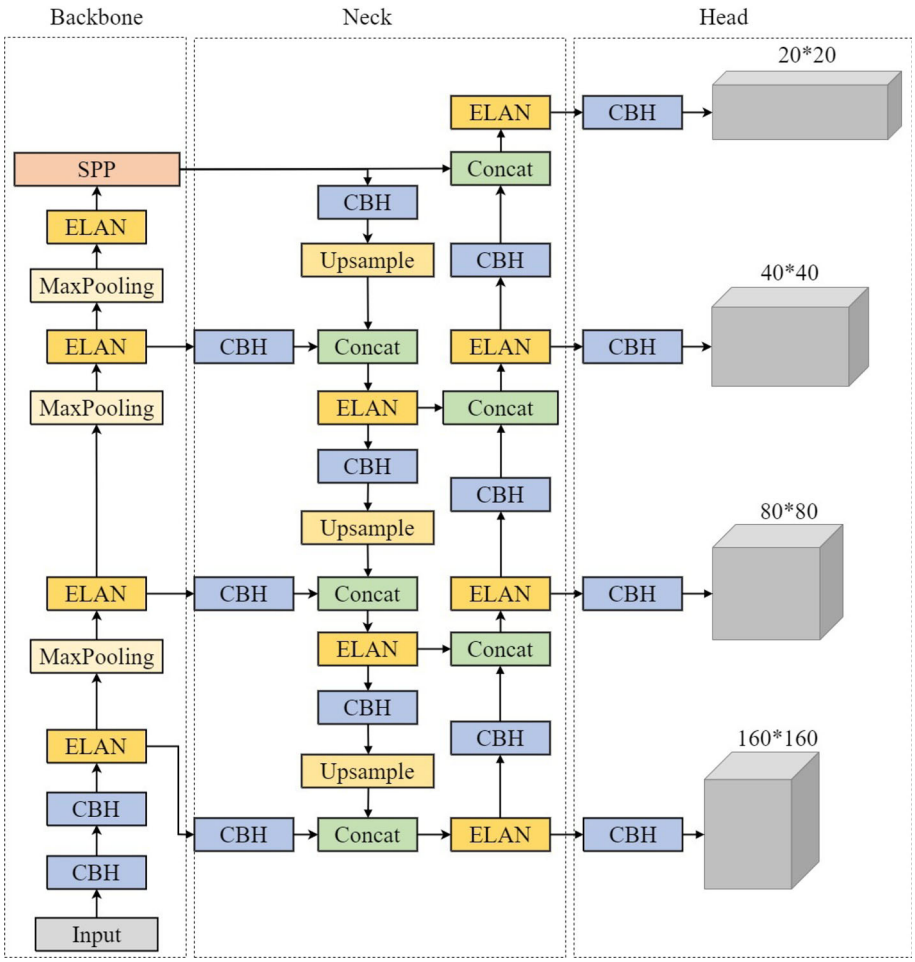


**Fig. 2** The HF-YOLO model structure mainly consists of three parts: backbone (feature extraction), neck (feature fusion), and head (detection)

The backbone component incorporates encapsulated convolutional blocks (CBH), MaxPooling, ELAN, and SPP modules for efficient feature extraction, as depicted in Fig. 3. CBH is defined as a sequence of operations:

$$X_{out} = HardSwish(BN(Conv(X_{in}))) \tag{1}$$

where $X_{in}$ and $X_{out}$ respectively represent the feature maps of the input and output. apply a $3 \times 3$ convolution operation to $X_{in}$, followed by batch normalization (BN), and subsequently an activation function (HardSwish).

Specifically, the ELAN module combines concepts from VoVNet [38] and CSPNet [39], utilizing a gradient path strategy to control the shortest and longest gradient paths in each layer. This enables different computational units to learn diverse information, maximizing parameter utilization efficiency. Additionally, the ELAN module ensures stable model learning by directly propagating information to update the weights of each computational unit, thereby mitigating degradation issues during training. Its gradient path design strategy pro-

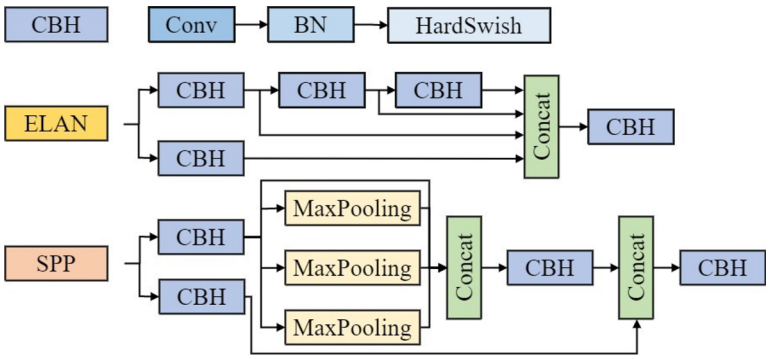


**Fig. 3** CBH consists of the operations convolution (Conv), batch normalization (BN), and the HardSwish activation function. The ELAN module employs a multi-branch structure, incorporating residual connections after each convolutional operation. This approach alleviates the issue of gradient vanishing associated with the increase in model depth. SPP module is overall composed of two parallel branch structures. The first part first undergoes CBH operation, then goes through max pooling with kernel sizes of $5\times5$, $9\times9$, and $13\times13$, followed by a residual connection. The second part consists of a residual connection with CBH operation. Finally, these two parts are concatenated to obtain the output

motes efficient parameter utilization, enabling the network to achieve higher accuracy without the need for additional complex architectures.

Spatial pyramid pooling (SPP) can generate fixed-size outputs, effectively addressing the issue of repetitive feature extraction in convolutional neural networks, while also reducing computational costs.

Following the backbone, the neck component focuses on feature fusion and information propagation. It integrates various techniques, such as feature concatenation, to combine features from different levels of the backbone. This facilitates the integration of low-level localization information and high-level semantic information, empowering the model to effectively handle pedestrians of diverse scales and occlusion levels.

The Head of the HF-YOLO model consists of four detection layers, which perform object detection on feature maps with sizes of $20 \times 20$, $40 \times 40$, $80 \times 80$, and $160 \times 160$, respectively. These detection layers enable the model to capture pedestrians at different scales and generate precise detection results.

By combining these components, the HF-YOLO model addresses the challenges of pedestrian detection by enhancing feature extraction, facilitating feature fusion, and enabling accurate detection across various scales.

### 3.2 Optimizing Feature Representation with the HardSwish Activation Function

In convolutional neural networks, the output of each layer is obtained by applying a linear transformation to the input from the previous layer. However, without an activation function, the network's output remains a linear combination of the inputs, regardless of its depth. Activation functions play a crucial role in introducing non-linearity to the data, enabling the network to capture complex patterns and effectively represent non-linear mappings between input and output domains. By incorporating non-linear transformations, activation functions enhance the expressive power of the network, facilitating the learning of intricate and abstract representations. This, in turn, enables the network to effectively model and extract features from high-dimensional data.

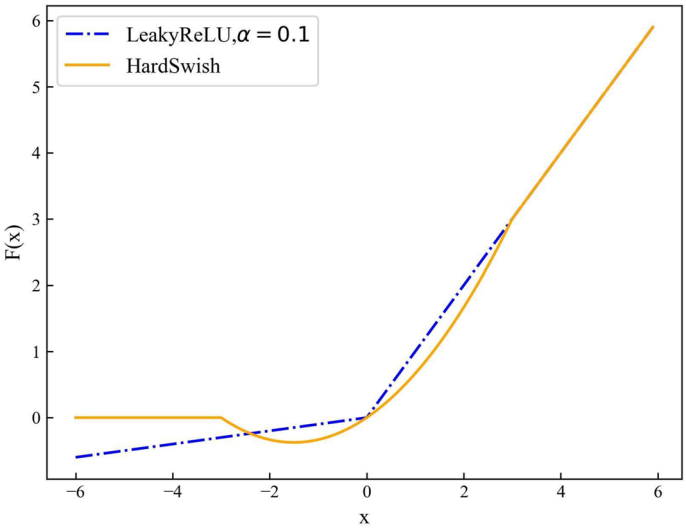


**Fig. 4** The plotting of the Leaky ReLU and HardSwish activation functions with it's non-monotonic bump for x less than 0

In model architectures, it is common to use Conv-BatchNorm-LeakyReLU (CBL) convolutional blocks, where convolutional operations (Conv), batch normalization (BN), and activation functions (LeakyReLU) are encapsulated together. The inclusion of activation functions introduces non-linearity, allowing the model to learn and represent intricate features. The LeakyReLU activation function is an adaptation of ReLU that introduces a small slope for negative values, addressing the issue of vanishing gradients encountered in ReLU:

$$LeakReLU(x) = \begin{cases} x, & x \geq 0 \\ 0.1x, & x < 0 \end{cases} \tag{2}$$

However, due to its near-linear behavior, LeakyReLU may not be optimal for detecting complex patterns in intricate data. The HardSwish activation function is defined as:

$$HardSwish(x) = \begin{cases} 0, & x \leq -3 \\ x, & x \geq 3 \\ \frac{x(x+3)}{6}, & otherwise \end{cases} \tag{3}$$

The most prominent distinction between Hard-Swish and Leaky ReLU is the non-monotonic concavity when x is less than 0. As shown in Fig. 4, the HardSwish [40] activation function undergoes a smoothing process, which enhances its smoothness and continuity. This characteristic facilitates gradient computation and optimization, mitigating problems such as gradient explosion and vanishing gradients. Consequently, it contributes to an accelerated convergence rate of the model. Compared to LeakyReLU, HardSwish introduces a stronger non-linearity while maintaining computational efficiency, thereby amplifying the representational capacity of neural networks. Consequently, replacing the activation function with HardSwish leads to the construction of the Conv-BatchNorm-HardSwish (CBH) convolutional block, which is used to build the ELAN and SPP modules.

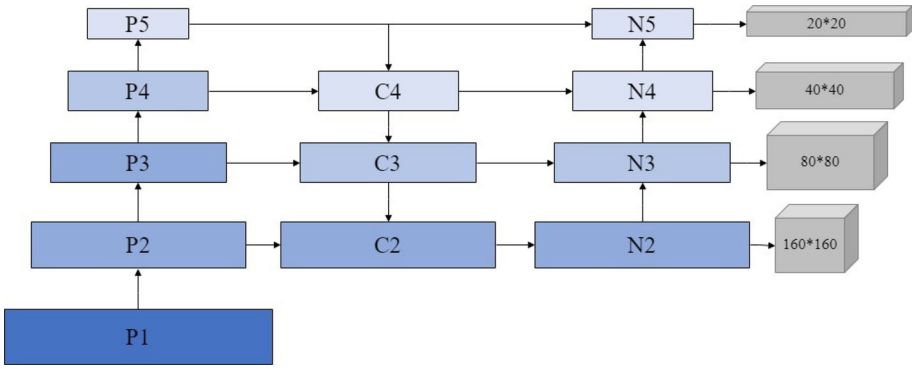


**Fig. 5** The features of {P2,P3,P4,P5} extracted from the backbone were fused from top to bottom

### 3.3 Fusion of High-Resolution Features

The detection of pedestrians in an image presents challenges due to the inherent variability in their distances from the camera, resulting in diverse scales. This scale variability can lead to scale imbalance and impact the performance of detection algorithms. Additionally, occluded pedestrians often lack discriminative features, which can result in false positives or false negatives during the detection process. Detection algorithms leverage shallower feature maps for precise spatial information and deeper feature maps for semantic context. By fusing feature maps from different layers, it becomes possible to exploit the complementary characteristics of both shallow and deep features. This fusion strategy enables the algorithm to capture richer contextual information, achieve multi-scale receptive fields, enhance detection capacity for objects of varying scales and occluded targets, and ultimately improve object localization precision.

During the feature extraction process, successive down-sampling operations are employed to derive informative feature representations. As the feature map's resolution decreases, it encapsulates higher-level semantic information pertinent to the targets. However, challenges arise, particularly in scenarios involving occluded pedestrians, where multiple individuals might share a common feature representation. Consequently, this can lead to the exclusion of occluded targets, hampering detection accuracy.

To bolster HF-YOLO's detection performance in handling diverse object scales and occluded pedestrians, we enhance the feature fusion module, illustrated in Fig. 5. Initially, we augment the C3 layer features using operations like convolution and upsampling. These augmented features are then fused with the P2 layer features extracted from the backbone network, leveraging the Concatenation (Concat) operation. This fusion yields C2, C3, C4, with its formulation outlined as follows:

$$C4 = Cat[U(CBH(P5)), CBH(P4)] \tag{4}$$

$$C3 = Cat[U(CBH(E(C4)), CBH(P3)] \tag{5}$$

$$C2 = Cat[U(CBH(E(C3)), CBH(P2)] \tag{6}$$

where $Cat$ concatenates the feature maps along the channel dimension, $CBH$ represents the Conv_BN_HardSwish operation, adjusting the number of channels. $U$ enlarges the feature map to twice its original size. $E$ represents the feature extraction operation. This fusion facilitates the propagation of high-level semantic information.

Furthermore, we employ a bottom-up fusion process to propagate the localization information from the lower layers to the higher layers. This process generates N2, N3, N4, N5, where the features from the lower layers carry important positional details. Its expression is as follows:

$$N2 = E(C2) \tag{7}$$

$$N3 = E(Cat[CBH(N2), C3]) \tag{8}$$

$$N4 = E(Cat[CBH(N3), C3]) \tag{9}$$

$$N5 = E(Cat[CBH(N4), P5]) \tag{10}$$

Following two iterations of feature fusion operations, the amalgamation of high-level semantic information and low-level localization details serves to bolster the efficacy of feature extraction for occluded targets.

Based on the fused feature maps, we construct a detection layer with a dimension of 160 $\times$ 160 to enhance the detection capability, particularly for small-sized targets. This increase in spatial resolution aids in capturing finer details and improving localization accuracy. The fusion process described above is visually depicted in Fig. 5.

### 3.4 Bounding Box Regression Loss Function

The bounding box regression loss function plays a crucial role in object detection algorithms as it encompasses accurate localization of objects. By comparing the predicted bounding boxes with the ground truth boxes, the bounding box regression loss value is computed, allowing the model to continuously optimize and improve the accuracy of object localization.

In object detection models, the complete IoU (CIoU) [41]metric is widely utilized for computing regression losses. The CIoU metric incorporates three essential geometric factors—the overlapping area, distance, and aspect ratio of the bounding box—into the loss calculation. This comprehensive metric provides a more accurate measure of the spatial discrepancy between predicted and ground truth bounding boxes. By considering these geometric factors, CIoU facilitates a more precise evaluation of the localization accuracy, enabling the optimization of object detection models for improved precision and robustness.

The formula for CIoU is as follows:

$$L_{CIoU} = 1 - IoU + \left( \frac{\rho^2 \left(b, b^{gt}\right)}{c^2} + \alpha v \right) \tag{11}$$

Where IoU represents the intersection over union of bounding boxes A and B, which is calculated using the following formula:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{12}$$

where $b$ represents the predicted bounding box's center coordinates, $b^{gt}$ represents the ground truth bounding box's center coordinates, $\rho^2(b, b^{gt})$ represents the squared distance between the center points of the two boxes, $c^2$ represents the squared diagonal length of the minimum enclosing rectangle of the two boxes. $v$ is a parameter that measures the aspect ratio, $w$ and $h$ represent the width and height of the predicted box, and $h^{gt}$ and $h^{gt}$ represent the width and height of the ground truth box, defined as follows:

$$v = \frac{4}{\pi^2} \left( \arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \tag{13}$$

$\alpha$ is a weight parameter defined as follows:

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{14}$$

In comparison to other regression loss functions, CIoU loss has demonstrated notable advancements in terms of convergence speed and detection accuracy. However, one inherent limitation of CIoU loss is its neglect of the inherent imbalance within the regression samples. During the training process, there is often a scarcity of high-quality samples accompanied by an abundance of low-quality samples. As a result, the contribution of gradients derived from the former category to the overall regression gradient is diminished. To overcome this challenge and further enhance the precision of pedestrian target localization, we propose the utilization of F-CIoU.

F-CIoU introduces the IoU as a balancing factor to weigh the CIoU loss, as exemplified by Equation 5, where $\beta$ is assigned a value of 0.5. By incorporating F-CIoU, higher-quality samples with larger IoU values incur greater losses and correspondingly higher weights. Consequently, the model places increased emphasis on the regression of high-quality samples, thereby promoting superior localization accuracy. The utilization of F-CIoU addresses the issue of imbalance in the regression samples, enabling the model to better optimize its performance by giving appropriate attention to different sample categories based on their quality.

$$L_{F-CIoU} = IoU^{\beta}\left(1 - IoU + \left(\frac{\rho^2\left(b, b^{gt}\right)}{c^2} + av\right)\right) \tag{15}$$

Following the aforementioned enhancements, we have derived the HF-YOLO model. The HF-YOLO model integrates the proposed improvements, including the novel feature fusion module, the utilization of the HardSwish activation function, and the introduction of the F-CIoU loss function. These enhancements collectively contribute to the improved performance of the HF-YOLO model in handling objects of various scales and occluded pedestrians, ultimately leading to enhanced object detection accuracy.

## 4 Experiment

In this section, we will provide an overview of the dataset used and the experimental setup employed in our study. We will then introduce the evaluation metrics utilized to assess the performance of the proposed HF-YOLO model. Additionally, we will present the experimental evaluations conducted on activation functions and the bounding box regression loss function. Subsequently, we will compare the results obtained from our proposed model with existing models through comparative analysis. Finally, we will discuss the ablation experiments performed to investigate the individual contributions of different components in the HF-YOLO model.

### 4.1 Experimental Setup

The experimental setup in this study is described in Table 1. In this study, we utilized the SGD optimizer for training the HF-YOLO model. The model was configured with an input image size of 640×640 pixels. The batch size is 64. The initial learning rate was set to 0.001,, and the cosine annealing algorithm is used to update the learning rate. The momentum factor

**Table 1** Experimental environment

| Name | Version |
|---|---|
| Operating System | Ubuntu20.0 |
| CPU | Intel® Xeon® Glod 5230 CPU @2.20 GHz |
| GPU | NVIDIA GeForce RTX A4000 |
| Python | 3.8 |
| Pytorch | 1.11.0 |
| CUDA | CUDA11.3 |

of 0.9 was employed during optimization. The total training duration comprised 300 epochs, during which the model was iteratively updated and fine-tuned using the training data.

### 4.2 Datasets

The dataset used in this study consisted of two main sources: the CrowdHuman dataset and the WiderPerson dataset.

The CrowdHuman dataset [42] is composed of images primarily obtained from Google search. It comprises complex scenes with dense pedestrian crowds. The training set of this dataset contains approximately 470,000 instances, with an average of around 23 people per image. The annotation information includes three types: full body, visual body, and head. For this study, the full body annotations were selected for training the model.

The WiderPerson dataset [43] is specifically designed for outdoor pedestrian detection. It consists of 13,382 images, and the annotations include five categories: pedestrian, cyclist, partially visible person, crowd, and ignore region. In this study, the first three categories (pedestrian, cyclist, partially visible person) were combined and treated as the "person" class.

To create the dataset for our experiments, we extracted a total of 16,000 images from the CrowdHuman and WiderPerson datasets. These images were then partitioned into three subsets: a training set, a validation set, and a testing set, following an 8:1:1 ratio. All experimental results were obtained in the test set.

### 4.3 Evaluation Metrics

The performance evaluation of the model in this study encompasses four key metrics: precision, recall, mean Average Precision (mAP), and detection time.

Precision quantifies the accuracy of positive predictions, serving as a measure of false detections. It is computed using the formula:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

where TP represents the number of correctly detected objects by the model, and FP denotes the count of objects falsely detected by the model.

Recall, on the other hand, gauges the probability of identifying positive samples within the predicted results, indicating the extent of missed detections. The calculation is given by:

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

**Table 2** Comparison of experimental results of activation function

| Activation function | mAP@0.5 (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| SiLU | 77.83 | 85.55 | 65.79 |
| GELU | 77.93 | 83.73 | 66.14 |
| LeakyReLU | 78.08 | 84.90 | **66.88** |
| HardSwish | **78.70** | **86.59** | 66.03 |

Bold font is designated to signify the optimal result pertaining to the respective evaluation metric

where FN represents the number of positive samples overlooked by the model.

The average precision (AP) is determined by computing the area under the precision-recall curve, which provides an assessment of the detection quality for each class. The mean average precision (mAP) corresponds to the average AP values across multiple classes and is derived as follows:

$$AP = \int_0^1 Pecision(Recall) d_{Recall} \tag{18}$$

$$mAP = \frac{1}{m} \sum_{i=1}^{m} AP_i \tag{19}$$

## 4.4 Experimental Results and Analysis

### 4.4.1 Experimental Comparison of Activation Functions

To ascertain the effectiveness of the HardSwish activation function, a comparative evaluation was performed, contrasting it with alternative activation functions, including SiLU [44], GELU [45], and LeakyReLU. The results of the experiments, presented in Table 2, unequivocally establish the superiority of HardSwish in terms of performance. These findings provide compelling evidence of its ability to introduce a higher level of non-linearity to the model, thereby augmenting its capacity to represent intricate patterns and complex relationships in the data.

### 4.4.2 Comparative Experiment of Bounding Bbox Regression Loss Function

To assess the effectiveness of the F-CIoU loss function, a comparative analysis was conducted, comparing it with alternative loss functions, including CIoU, SIoU [46], and WiseIoU [47]. The experimental results, presented in Table 3, yield valuable insights.

The SIoU loss incorporates an additional angle loss component to address the issue of predicted boxes exhibiting undesired wandering behavior during training. As a result, it achieves a modest improvement of 0.41% in recall compared to the CIoU loss. However, this improvement comes at the cost of a slight decline in both average precision and precision measures.

The WiseIoU loss, which introduces an attention-based regression loss, aims to enhance object detection performance. However, when compared to the CIoU loss, it does not yield any substantial improvement in model performance.

In contrast, the F-CIoU loss, by incorporating an IoU weighting factor, effectively mitigates the issue of regression imbalance. It outperforms the original loss function by achieving

**Table 3** Comparative experimental results of boundary box regression loss function

| Loss function | mAP@0.5 (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| CIoU | 80.72 | 86.21 | 68.59 |
| SIoU | 80.57 | 85.66 | 69.00 |
| WiseIoU | 80.12 | 85.45 | 67.77 |
| F-CIoU | **81.60** | **86.25** | **71.71** |

Bold font is designated to signify the optimal result pertaining to the respective evaluation metric

**Table 4** Experimental results in comparison with other models

| Model | mAP@0.5(%) | Precision(%) | Recall(%) | Times(ms) |
|---|---|---|---|---|
| SSD | 65.30 | 72.24 | 66.99 | 13.0 |
| RTDETR | 80.10 | 83.06 | 69.09 | 22.9 |
| YOLOv3tiny | 67.40 | 75.80 | 58.30 | **3.3** |
| YOLOv4tiny | 70.93 | 82.49 | 61.53 | 4.1 |
| YOLOv5s | 79.50 | 85.45 | 67.23 | 8.0 |
| YOLOv6n | 79.45 | 84.43 | 68.86 | 7.0 |
| YOLOv7tiny | 78.08 | 84.90 | 66.88 | 7.7 |
| YOLOv8n | 79.78 | 84.81 | 68.69 | 7.0 |
| Ours | **81.60** | **86.25** | **71.71** | 8.5 |

Bold font is designated to signify the optimal result pertaining to the respective evaluation metric

a notable improvement of 0.88% in mean average precision (mAP) and a significant 3.12% increase in recall. These results clearly demonstrate the effectiveness of the F-CIoU loss in optimizing the model's performance by addressing the regression imbalance and improving detection accuracy.

### 4.4.3 Experimental Results in Comparison with Other Models

The HF-YOLO model, incorporating the proposed enhancements, was evaluated against several state-of-the-art detection models, including SSD, RTDETR, YOLOv3tiny, YOLOv4tiny, YOLOv5s, YOLOv6n, YOLOv7tiny, and YOLOv8n, in a series of comparative experiments.

Table 4. presents compelling evidence supporting the superiority of our HF-YOLO model across various crucial metrics such as mean average precision (mAP), precision, and recall. The results showcase a clear advancement compared to all other models examined. Specifically, HF-YOLO demonstrates superior performance over SSD and RTDETR, registering an impressive increase of 16.3% in detection accuracy compared to SSD and 1.5% compared to RTDETR. Moreover, our model achieves a reduction in detection time by 4.5ms in contrast to SSD and 14.4ms compared to RTDETR.

Although YOLOv3tiny and YOLOv4tiny exhibit the quickest detection times, they compromise on mAP, precision, and recall metrics. Meanwhile, our HF-YOLO model, while maintaining a comparable detection time to YOLOv5s, YOLOv6s, YOLOv7tiny, and YOLOv8n, significantly outperforms them across all other evaluated metrics. These outcomes underscore the practical applicability and efficacy of our proposed model in real-world object detection scenarios.

**Table 5** Ablation experiment design

| Model | HardSwish | Detection Layer | F-CIoU |
|---|---|---|---|
| YOLOv7tiny | ✗ | ✗ | ✗ |
| ① | ✓ | ✗ | ✗ |
| ② | ✓ | ✓ | ✗ |
| ③ | ✓ | ✓ | ✓ |

**Table 6** Results of ablation experiment

| Model | mAP@0.5 (%) | Precision (%) | Recall (%) | Times (ms) |
|---|---|---|---|---|
| YOLOv7tiny | 78.08 | 84.90 | 66.88 | 7.7 |
| ① | 78.70 | **86.59** | 66.03 | **7.7** |
| ② | 80.72 | 86.21 | 68.59 | 8.5 |
| ③ | **81.60** | 86.25 | **71.71** | 8.5 |

Bold font is designated to signify the optimal result pertaining to the respective evaluation metric

### 4.4.4 Results and Analysis of Ablation Experiments

To evaluate the efficacy of each module improvement, a series of ablation experiments were conducted to analyze the impact of different modules. The ablation experiment design, outlined in Table 5, utilized YOLOv7tiny as the baseline model. The following improvements were investigated:

- Improvement①: Replacement of LeakyReLU with HardSwish activation function.
- Improvement②: Fusion of high-resolution features and addition of small object detection layer.
- Improvement③: Modification of regression loss function from CIoU to F-CIoU.

By conducting these ablation experiments and comparing the results with the baseline model, the effectiveness of each improvement can be assessed.

The experimental results, presented in Table 6, demonstrate the outcomes of the conducted ablation experiments.

In experiment①, involving the replacement of the activation function, the introduction of HardSwish resulted in increased non-linearity, leading to improved feature representation capabilities. Compared to the original model, there was a noticeable enhancement of 0.62% in mean average precision (mAP) and 1.69% in Precision. The detection time remained consistent at 7.7ms, indicating the effectiveness of the HardSwish activation function in improving model performance.

Building upon the findings of Experiment①, Experiment② introduced additional detection layers, enhancing the model's ability to detect pedestrians of varying scales. This led to a significant improvement of 2.64% in mAP and 1.71% in Recall, with the detection time increasing by 0.8ms to reach 8.5ms, compared to the original model.

Experiment③ focused on improving the regression loss function. By doing so, the detection rate of high-quality samples was enhanced, resulting in a substantial increase in Recall to 71.71%. The mAP was elevated to 81.60%, and the Precision reached 86.25%. The detection time remained consistent at 8.5 ms.

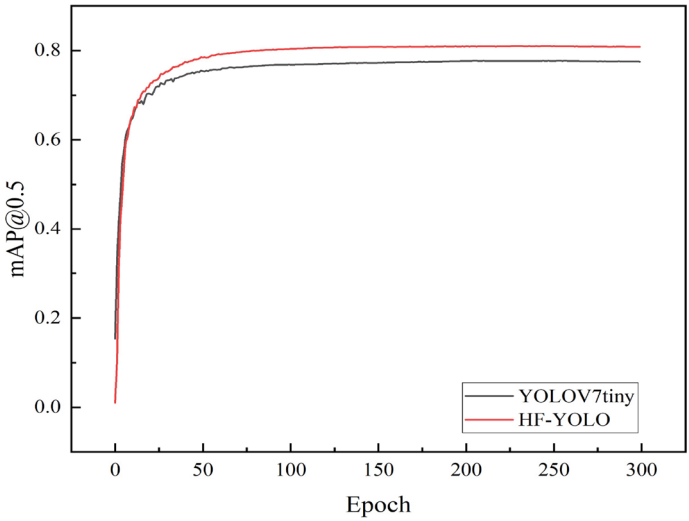


**Fig. 6** Comparison of mAP between HF-YOLO and YOLOv7tiny

The mAP comparison between HF-YOLO and YOLOv7tiny is presented in Fig. 6, where it can be observed that HF-YOLO exhibits higher accuracy, achieving an improvement of 3.52 in mAP.

### 4.4.5 Visualization of Detection Results

Figure 7 serves as a visual validation of the proposed algorithm's efficacy, displaying detection results for SSD, RTDETR, YOLOv3tiny, YOLOv4tiny, YOLOv5s, YOLOv6s, YOLOV7tiny, YOLOv8n, and HF-YOLO algorithms. The purpose of this experiment is to showcase the superior performance of the HF-YOLO algorithm.

The results indicates a significant difference between our proposed model and others. In the first set of images, SSD, YOLOv3tiny, YOLOv6s, and YOLOv8n display conspicuous shortcomings in detecting small targets. Moreover, excluding our improved algorithm, other detection outcomes suffer from imprecise localization. Moving to the second and third sets of images, depicting more complex scenes with partially occluded pedestrians, our algorithm consistently exhibits commendable detection capabilities.

Relative to the results of YOLOv7tiny, our enhanced algorithm notably diminishes false positives and elevates the accuracy of target localization. Furthermore, it consistently achieves precise target detection even in complex scenes.

## 5 Conclusion

In conclusion, this paper presents HF-YOLO, an advanced pedestrian detection model that effectively addresses the challenges associated with pedestrian detection, including scale variations and occlusion. By leveraging feature fusion from shallow and deep layers, HF-YOLO enhances the model's detection performance, resulting in improved accuracy and robustness. Moreover, to tackle the issue of imbalance in bounding box regression, HF-YOLO incorporates a balancing factor that prioritizes the accurate localization of high-quality samples. This

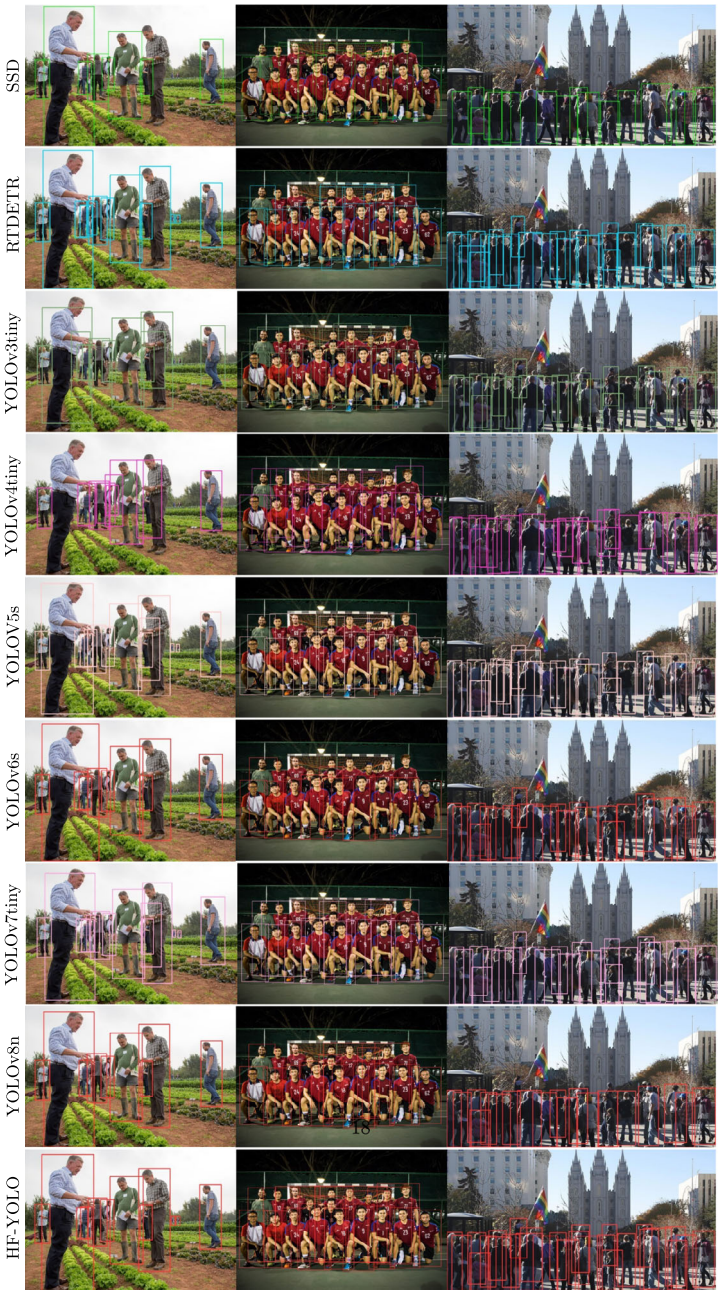


**Fig. 7** In contrast to the detection outcomes of SSD, RTDETR, and other algorithms within the YOLO series, the HF-YOLO algorithm consistently demonstrates accurate detection of pedestrian targets, exhibiting proficiency in scenarios characterized by both small-scale targets and occlusions

ensures that the model focuses on optimizing the detection of relevant objects and improves overall detection efficacy. Experimental evaluations conducted in this study provide compelling evidence of the effectiveness of the proposed HF-YOLO model. It outperforms the baseline model, demonstrating significant improvements in detection accuracy and overall performance. The results highlight the potential of HF-YOLO as an advanced pedestrian detection solution, with practical applications in real-world scenarios.



## References

1. Maqsood M, Yasmin S, Gillani S et al (2023) An efficient deep learning-assisted person re-identification solution for intelligent video surveillance in smart cities. Front Comp Sci 17(4):174329
2. El Hamdani S, Benamar N, Younis M (2020) Pedestrian support in intelligent transportation systems: challenges, solutions and open issues. Transp Res part C Emerg Technol 121:102856. https://doi.org/10.1016/j.trc.2020.102856
3. Lee S, Lee S, Seong H et al (2023) Fallen person detection for autonomous driving. Expert Syst Appl 213:119242. https://doi.org/10.1016/j.eswa.2022.119242
4. Wang K, Li G, Chen J et al (2020) The adaptability and challenges of autonomous vehicles to pedestrians in urban china. Accid Anal Prev 145:105692. https://doi.org/10.1016/j.aap.2020.105692
5. Hariyono J, Jo KH (2017) Detection of pedestrian crossing road: a study on pedestrian pose recognition. Neurocomputing 234:144–153. https://doi.org/10.1016/j.neucom.2016.12.050
6. Lee S, Rim J, Jeong B, et al (2023) Human pose estimation in extremely low-light conditions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 704–714
7. Wong PKY, Luo H, Wang M et al (2021) Recognition of pedestrian trajectories and attributes with computer vision and deep learning techniques. Adv Eng Inf 49:101356. https://doi.org/10.1016/j.aei.2021.101356
8. Feng J, Wu A, Zheng WS (2023) Shape-erased feature learning for visible-infrared person re-identification. In: 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, pp 22752–22761
9. Paul V, Michael J (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, pp I–I
10. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), pp 886–893
11. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110. https://doi.org/10.1016/j.infrared.2021.103694
12. Dai X, Hu J, Zhang H et al (2021) Multi-task faster R-CNN for nighttime pedestrian detection and distance estimation. Infrared Phys Technol 115:103694. https://doi.org/10.1016/j.infrared.2021.103694
13. Xue Y, Ju Z, Li Y et al (2021) MAF-YOLO: multi-modal attention fusion based yolo for pedestrian detection. Infrared Phys Technol 118:103906. https://doi.org/10.1016/j.infrared.2021.103906
14. Jain DK, Zhao X, González-Almagro G et al (2023) Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. Inf Fusion 95:401–414
15. Girshick R, Donahue J, Darrell T, et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
16. Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448

17. Ren S, He K, Girshick RB et al (2015) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149. https://doi.org/10.1109/tpami.2016.2577031
18. Liu W, Anguelov D, Erhan D, et al (2016) Ssd: Single shot multibox detector. In: Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14, pp 21–37
19. Redmon J, Divvala S, Girshick R, et al (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
20. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
21. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. Preprint at https://arXiv.org/abs/1804.02767
22. Zhou X, Wang D, Krähenbühl P (2019) Objects as points. Preprint at https://arXiv.org/abs/1904.07850
23. Tian Z, Shen C, Chen H et al (2020) FCOS: a simple and strong anchor-free object detector. IEEE Trans Pattern Anal Mach Intell 44(4):1922–1933
24. Yu J, Rui Y, Tao D (2014) Click prediction for web image reranking using multimodal sparse coding. IEEE Trans Image Process 23(5):2019–2032
25. Cao J, Cholakkal H, Anwer RM, et al (2020) D2det: Towards high quality object detection and instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11485–11494
26. Yu J, Tan M, Zhang H et al (2019) Hierarchical deep click feature prediction for fine-grained image recognition. IEEE Trans Pattern Anal Mach Intell 44(2):563–578
27. Woo S, Park J, Lee JY, et al (2018) Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
28. Lv W, Xu S, Zhao Y, et al (2023) Detrs beat yolos on real-time object detection. Preprint at https://arxiv.org/abs/2304.08069
29. Zhang L, Lin L, Liang X, et al (2016) Is faster R-CNN doing well for pedestrian detection? In: Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14, Springer, pp 443–457
30. Zhang S, Wen L, Bian X, et al (2018) Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In: Proceedings of the European conference on computer vision (ECCV), pp 637–653
31. Liu S, Huang D, Wang Y (2019) Adaptive NMS: refining pedestrian detection in a crowd. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6459–6468
32. Chu X, Zheng A, Zhang X, et al (2020) Detection in crowded scenes: one proposal, multiple predictions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12214–12223
33. Xia H, Ma J, Ou J et al (2021) Pedestrian detection algorithm based on multi-scale feature extraction and attention feature fusion. Digital Signal Process 121:103311. https://doi.org/10.1016/j.dsp.2021.103311
34. Li Q, Qiang H, Li J (2021) Conditional random fields as message passing mechanism in anchor-free network for multi-scale pedestrian detection. Inf Sci 550:1–12. https://doi.org/10.1016/j.ins.2020.10.049
35. Wang M, Ma H, Liu S et al (2023) A novel small-scale pedestrian detection method base on residual block group of CenterNet. Comput Stand Interfaces 84:103702. https://doi.org/10.1016/j.csi.2022.103702
36. Gao S, Cheng MM, Zhao K et al (2019) Res2Net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662
37. Wang CY, Bochkovskiy A, Liao HYM (2023) 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, pp 7464–7475
38. Lee Y, won Hwang J, Lee S, et al (2019) An energy and GPU-computation efficient backbone network for real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 0–0
39. Wang CY, Liao HYM, Wu YH, et al (2020) 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, pp 390–391
40. Howard A, Sandler M, Chu G, et al (2019) Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1314–1324
41. Zheng Z, Wang P, Ren D et al (2021) Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans Cybern 52(8):8574–8586
42. Shao S, Zhao Z, Li B, et al (2018) Crowdhuman: a benchmark for detecting human in a crowd. Preprint at https://arXiv.org/abs/1805.00123
43. Zhang S, Xie Y, Wan J et al (2019) Widerperson: a diverse dataset for dense pedestrian detection in the wild. IEEE Trans Multimed 22(2):380–393

44. Elfwing S, Uchibe E, Doya K (2018) Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Netw 107:3–11. https://doi.org/10.1016/j.neunet.2017.12.012
45. Hendrycks D, Gimpel. K (2016) Bridging nonlinearities and stochastic regularizers with gaussian error linear units. Preprint at https://arxiv.org/abs/1606.08415
46. Gevorgyan Z (2022) Siou Loss: More powerful learning for bounding box regression. Preprint at https://arxiv.org/abs/2205.12740
47. Tong Z, Chen Y, Xu Z, et al (2023) Wise-iou: Bounding box regression loss with dynamic focusing mechanism. Preprint at https://arxiv.org/abs/2301.10051