sciendo

JAISCR, 2026, Vol. 16, No. 3, pp. 219 – 235
10.2478/jaiscr-2026-0011

AN IMPROVED YOLO ALGORITHM BASED ON CONCISE DECOUPLED HEAD FOR REAL-TIME OBJECT DETECTION IN NIGHT SCENARIOS

Yanhua Ma[1], Ke Lv[2], Li-Juan Liu[3,*], Hamid Reza Karimi[4]

[1] *School of Integrated Circuits, Dalian University of Technology Liaoning, 116023, Dalian, China*

[2] *School of Railway Intelligent Engineering, Dalian Jiaotong University Liaoning, 116014, Dalian, China*

[3] *School of Information Science, Beijing Language and Culture University Beijing, 100083, Beijing, China*

[4] *Department of Mechanical Engineering, Politecnico di Milano 20156, Milan, Italy*

[*] *E-mail: liulijuan@blcu.edu.cn, mishu06@163.com*

Submitted: 30th June 2025; Accepted: 27th December 2025

Abstract

This paper proposes CDH-YOLO, an efficient, real-time pedestrian detection model for nighttime RGB images. Built on YOLOv5, CDH-YOLO incorporates structural reparameterization to optimize the backbone network and integrates convolutional block attention module to enhance feature representation. Transposed convolution replaces nearest neighbor interpolation for upsampling to preserve semantic information. A lightweight decoupled head addresses spatial misalignment between classification and regression tasks, while SIoU loss improves training convergence and localization accuracy. Experiments on the KAIST dataset demonstrate that CDH-YOLO achieves superior accuracy with real-time performance, significantly outperforming existing methods in nighttime pedestrian detection.

Keywords: Convolutional block attention module (CBAM), decoupled head, structure reparameterization, YOLO

1 Introduction

Being a core component of computer vision, object detection is ubiquitously utilized across various sectors including but not limited to industry, agriculture, healthcare and other fields, and it relies heavily on an extremely deep convolutional neural network to extract the most salient features. [1] Object detection usually identifies and locates the interested object in the image by determining the location in the dataset. Among object detection techniques, pedestrian detection has become a major research focus across various fields. As a

key component of autonomous driving, pedestrian detection assists drivers in accurately and quickly avoiding pedestrians during the driving process. Thus the traffic safety can be effectively improved. Due to the extensive study of pedestrian detection technology. However, because the night scene is affected by the lighting conditions, the contrast difference of the pictures is large, and the color information is small, The phenomenon of missed and erroneous detections often occurs, which leads to present popular visual detection algorithms not being able to perform well in detection tasks. Therefore the detection ability of pedestrian detection algorithm under low-light conditions has evolved into a pressing issue that requires an immediate solution.

In the past decade, improved multimodal detection algorithms built upon classical algorithms have achieved better detection results, such as the combination of infrared devices and radar and other devices with object detection algorithms. Zhang et al. [2] developed LDHD-Net, a lightweight infrared detection algorithm with advanced feature extraction and attention modules to improve UAV detection accuracy and reduce false alarms in complex scenes. Wei et al. [3] proposed a pedestrian detection approach combining an enhanced UNet with YOLO, integrating visible and infrared data to improve accuracy in challenging infrared conditions. Shang et al. [4] combine infrared cameras with millimeter-wave radar to enhance nighttime pedestrian detection. By aligning and processing the two data sources separately, radar data compensates for missed infrared detections, significantly improving detection success in night scenes. Xue et al. [5] propose a real-time method for pedestrian detection that utilizes multi-channel attention fusion, which extracts features from two patterns and then fuses them by a pattern-weighted fusion module. Liu et al. [6] proposed a YOLOv8-based steel surface defect detection method using lightweight convolution, integrating ScConv, SPPF, GSConv, VoV-GSCSP, and the Coordinate Attention module, achieving excellent results in nighttime detection. In addition, Liu et al. [7] introduced an improved YOLOv8 algorithm with a FasterNet backbone for defect detection on printed circuit board (PCB) surfaces. This approach enhances detection accuracy and efficiency, addressing the challenges faced by traditional methods in complex manufacturing environments.

However, although the previous literature solves part of the problem through improved multimodal target detection algorithms, these multimodal approaches either in the pre-data set shooting and labeling or hardware equipment can bring great cost. Thus these approaches can not be applied to conventional object detection scenarios. Moreover, due to the challenges of nighttime detection with conventional cameras, few studies have focused on this topic.

In recent years, attention mechanisms that imitate human perception by focusing on certain important areas one after another to understand a whole scene have been extensively studied to enable optimization of network models. Attention mechanisms have proven to be beneficial to illustrate enhancements in the performance of tasks. Various attention mechanisms have been developed, each with unique characteristics. The Squeeze-and-Excitation (SE) module [8],and [9] adaptively recalibrates channel-wise feature responses but lacks spatial attention. The Convolutional Block Attention Module (CBAM) [10] sequentially infers both channel and spatial attention maps, providing a more comprehensive attention mechanism. The Detection Transformer (DETR) [11] employs self-attention for end-to-end object detection but suffers from high computational complexity and slow convergence. Compared to SE and DETR, CBAM offers a better balance between performance and computational efficiency, making it more suitable for real-time applications. And then Cao et al. [12] propose an innovative attention mechanism for mobile networks through the incorporation of location data into channel attention, which is capable of effectively capturing the global context. And Ma et al. [13] proposed the Adaptive Attention Module (AAM), which integrates both channel and spatial attention submodules to effectively balance model performance and complexity. By utilizing an adaptive mecha-

nism and a group-interact-aggregate strategy, AAM enhances feature representation and emphasizes important regions, demonstrating significant improvements across various tasks and datasets. By preserving precise positional information in one dimension, this approach enables modeling long-range dependencies along the other spatial dimension.Liu et al. [14] further optimized the YOLOv8 architecture with FasterNet backbone, achieving efficient defect detection on printed circuit boards even in low-light industrial environments. Another study by Liu et al. [15] proposed a dynamic feature selection and fusion strategy, which provides valuable insights for handling complex background interference in nighttime detection tasks.

Inspired by previous work, this paper proposes the CDH-YOLO target detection algorithm, only by means of images taken by RGB cameras as training samples, by introducing attention module, decoupling classification task and regression task and other improvement methods, compensates for the deficiencies of the multi-modal detection approach and substantially enhances pedestrian detection performance in nighttime scenes. Although many domestic and foreign scholars have conducted research on pedestrians detection in nighttime scenes, due to poor lighting conditions and potential interference from the colors of streetlights, it is difficult to extract clear pedestrians features. Consequently, challenges exist in nighttime scenes including a lack of diverse datasets, high rates of missed detections for small target pedestrians that cannot meet real-time detection requirements. Results from experiments conducted on the KAIST dataset show that the proposed CDH-YOLO algorithm attains exceptional precision and rapid prediction rates for detecting pedestrians in low visibility nighttime scenes. The main innovations introduced found in this work are outlined in the following:

1. A novel reparameterization module is proposed to boost network performance and reduce parameter overhead, making the network model more lightweight and suitable for real-time detection applications. Integrated a low-cost decoupling head structure within the architecture to effectively mitigate spatial biases in pedestrian detection, achieving substantial computational savings.

2. Employed transposed convolutional upsampling, allowing the network to autonomously learn convolutional kernel weights. This adaptive learning mechanism enhances the network's capability to handle pedestrians of varying scales and poses, improving spatial resolution restoration and overall model representation.

3. Incorporated the CBAM attention module and the SIoU metric into the model architecture, improving the model's ability to discern pedestrians under varying lighting and background conditions, and achieving finer localization in complex nighttime scenes. This dual enhancement focuses on crucial pedestrian positions, significantly boosting detection accuracy and robustness.

The organization of the remaining sections is: Section 2 and Section 3 summarizes recent research relevant to object detection. In Section 4, we summarize the YOLOv5 object detection algorithm. In Section 5 we elaborate the details of the components in our proposed approach and present experimental validation. Finally, we conclude this algorithm study in Section 6.

2 Related Work

In modern research, object detection approaches are distinguished as operating in one detection step or two, known as single-stage and two-stage. The former refers to algorithms that do not independently display the extracted regional proposal, and directly from the input image, detecting the class and location of targets present in a single inference pass. The classical single-stage algorithms include OverFeat [16], Single Shot multibox-Detector (SSD) [17], You Only Look Once (YOLO) [18], etc.Since feature identification and localization occur in one step, single-stage detectors exhibit higher speeds for these tasks. The latter refers

to detection algorithms that have an independent display of the region proposal extraction process, among which the more classical ones are the Region-Convolutional Neural Network (R-CNN) [19],Fast Region-Convolutional Neural Network (Fast R-CNN) [20], Spatial Pyramid Pooling Network (SPPNet) [21] and Mask R-CNN [22].

2.1 Two-stage Approaches

For RCNN [19], the sliding window approach is replaced with selective search to address the problem of redundant windows and reduce time complexity. Although these modern object detection algorithms have achieved significant performance gains over traditional approaches, they still face several challenges.To address the limitations of R-CNN models, He et al. proposed SPPNet [21]. It eliminates the need for fixed-size image inputs, prevents information loss from cropping or warping images, and avoids redundant convolution computations, thus significantly improving accuracy and speed. Then Girshick et al. proposed Fast R-CNN [20], building upon SPPNet by introducing Region of Interest (ROI) pooling, which not only demonstrated improved accuracy and faster test speeds than SPPNet, but also has substantially improved training speed. Only one year later, Ren et al. proposed Faster R-CNN [23], and it mainly integrates feature extraction, bounding box regression and classification, and region segmentation into one model. This end-to-end approach for object detection yielded substantial gains in overall performance. While region proposal-based object detection models have achieved ever-improving accuracy and inference speeds, the computational complexity of generating regional proposals remains a bottleneck that precludes real-time performance.

2.2 One-stage Approaches

The one-stage algorithm omits the candidate frame generation phase and simultaneously executes feature extraction, object categorization, and box prediction inside one convolutional network architecture, which also makes the one-stage algorithm faster than the two-stage algorithm. A single neural network can directly and simultaneously obtain the location of an object and the class. It belongs to the corresponding confidence probability after a single evaluation. A year later, Redmon et al. [24] proposed YOLOv2 based on it.To address the issue of saturated nonlinear models struggling to converge, batch normalization operations were utilized, which helped stabilize the training and accelerate convergence for these models. Redmon et al. [25] then proposed YOLOv3, which used Darknet-53 as the network backbone. In addition, the Neck section of YOLOv3 used a Feature Pyramid Network (FPN) architecture to detect three feature maps of different sizes, improving the network's ability to detect small targets. Two years later, Bochkovskiy et al. [26] proposed YOLOv4, which chose CSPDarknet53 as the backbone network, while many applicable algorithms were incorporated into the model. Some techniques that have been proposed to improve network generalization include adding weighted residual connections, partial connections across stages, self-adversarial training, normalizing across small batches, and regularization through dropblocks. These tuning tools allowed the model to achieve the optimal experimental results at that time.The YOLOv5 [27] algorithm was later introduced, which leapfrogged over YOLOv4 to directly became the most advanced algorithm for object detection to date, surpassing the prior version. Although the YOLOv5 algorithm is widely used in industry, there is still much room for improvement in specific scenarios.YOLOv6 [28] was launched by Meituan and its main work is to better adapt to GPU devices, introducing the RepVGG structure. The design of backbone and neck utilizes hardware advantages such as the computing characteristics of the processor core, memory bandwidth, etc. for effective inference.YOLOv7 [29] is a sequel to the YOLOv4 team, mainly aimed at optimizing model structure reparameterization and dynamic label allocation issues. The concept of YOLOv7 detection algorithm is similar to YOLOv4 and YOLOv5. Its proposed model structure is reparameterized, utilizing "extension" and "composite scaling" methods to ef-

fectively utilize parameters and computational complexity.A new label allocation method has been proposed. YOLOv8 from Ultralytics is the most recent iteration of their YOLO object detector and image segmenter. Building upon prior YOLO versions, YOLOv8 introduces new enhancements and optimizations that further advance the state-of-the-art(SOTA) in object detection, achieving improved accuracy, higher velocity, and greater flexibility compared to its predecessors. Overall, YOLOv8 represents the cutting edge in object detection and image segmentation tool that provides the latest SOTA technology and the ability to use and compare all previous YOLO versions, making it a choice that combines advantages.

2.3 Other Improvement Methods

In addition to the evolution of YOLO series detectors, recent research has also focused on optimizing network architectures and sampling strategies to improve detection performance, especially for challenging scenarios such as small object detection and complex backgrounds.

For aerial image small object detection, Zhang et al. [30] proposed Sinextnet, a novel model based on PP-YOLOE. Sinextnet incorporates a lightweight feature enhancement module and optimized neck structure to improve the detection accuracy of small objects in aerial images. This work demonstrates the importance of specialized architecture design for handling scale variations and low-resolution targets, which aligns with our focus on nighttime pedestrian detection where targets often appear small and blurry.

In terms of feature extraction and downsampling, Ji et al. [31] introduced Adaptive Separation Fusion (ASF), a novel downsampling approach in CNNs. ASF adaptively separates features into multiple groups and applies different downsampling strategies before fusing them, preserving more detailed information while reducing spatial dimensions. This innovative approach to feature preservation during downsampling provides insights complementary to our use of transposed convolution for upsampling, together forming a more comprehensive perspective on feature resolution maintenance throughout the network architecture.

These recent advancements in specialized detection architectures and feature processing methods further validate the importance of tailored network designs for specific detection challenges, supporting our approach of integrating multiple optimization strategies for nighttime pedestrian detection.

3 Basic Theory

Ultralytics proposed the YOLOv5 algorithm in 2020, and after continuous iteration, the highest version has now reached v7.1. According to depth and width, it can be decomposed into five distinct model species-YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5n, and YOLOv5x. This article selects the most commonly used and lightweight version of YOLOv5s v6.1 as the optimization baseline. The architecture of YOLOv5 v6.1, as shown in Figure 1 can be summarized by three central modules: Backbone, Neck and Head. The Backbone network of YOLOv5 is composed of Conv module, Cross Stage Partial (CSP) Bottleneck with 3 convolutions (C3) module and Spatial Pyramid Pooling-Fast (SPPF) module. The C3 module BottleNeck structures in the Backbone and Neck parts are different. The Neck utilizes FPN [32] and Path Aggregation Networks (PAN) [33] to improve the network's capability to integrate features for detecting objects of various sizes. The Head leverages the fused multi-scale features from the Neck to perform detection across multiple scales, improving detection capabilities for targets of varying sizes.

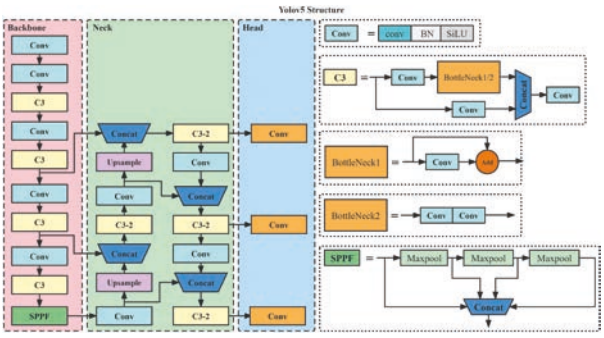

Figure 1. The YOLOv5 v6.1 Structure.

4 Model Improvement and Optimization

4.1 Backbone Improvements

4.1.1 CBAM Attention Module

The essential role of attention in deep learning has been thoroughly explored and validated by substantial prior research,see for instance [34–41]. To tackle the issue of high misdetection rates under low visibility nighttime conditions,we add an attention mechanism to enable concentrating on the most salient features while ignoring unnecessary ones. Thus, the model characterization capability can be increased. As shown in Figure 2. displays, this article incorporates the original C3 module of the CBAM [10] attention mechanism into the YOLOv5 backbone network, which includes two independent submodules, as shown in Figure 3, the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). All C3 modules in the backbone are replaced with C3CBAM to enhance feature representation. Since CBAM extracts information features by fusing channel feature information and spatial feature information, it can enhance the characterization information of low-visibility targets from two dimensions. Ablation Experiments show that CBAM substantially increases the detection effect of low-visibility targets.

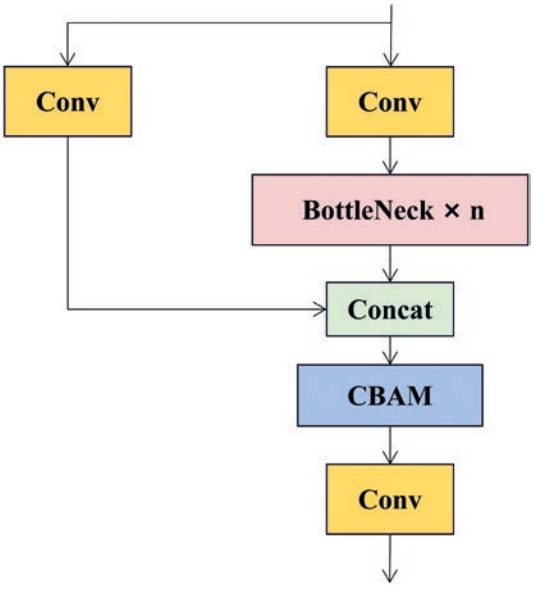

Figure 2. The C3CBAM Block Structure.

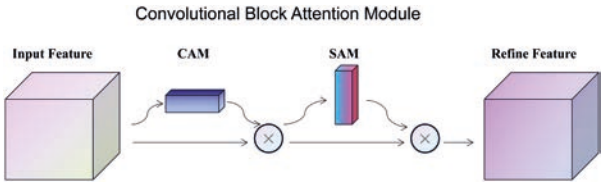

Figure 3. The CBAM Structure.

The CBAM operates in two steps: CAM and SAM. It takes an intermediate feature map $F \in R^{C\times H\times W}$ as input. First, to obtain two one-dimensional vectors, the input channel introduces a global maximum pooling layer and a global average pooling layer. These vectors are transmitted to the fully connected layer and added together to generate the one-dimensional channel attention vector $M_c \in R^{C\times 1\times 1}$, performed a multiplication operation with the input feature map to obtain the feature map F'. Second, the F' undergoes global maximum pooling layer and a global average pooling layer along the spatial dimensions to produce two 2D vectors. These vectors are concatenated and convolved to finally generate a 2D spatial attention map $M_s \in R^{1\times H\times W}$ is multiplied with F' element-wise to attend to salient spatial locations, as depicted in Figure 2. The process of CBAM attention algorithm can be described as follows:

$$F' = M_c(F)F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

where $\otimes$ symbol represents element multiplication. Before performing the multiplication operation, the vector based on channel attention and the vector based on spatial attention will match the input feature map's source space dimension and source channel dimension, respectively.The equation based on channel attention is

$$\begin{aligned} M_c(F) = \sigma\big(&MLP(AvgPool(F)) \\ &+ MLP(MaxPool(F))\big) \end{aligned} \tag{3}$$

The equation based on spatial attention is

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times 7}([AvgPool(F), MaxPool(F)])) \\ &= \sigma(f^{7\times 7}([F^s_{avg}, F^s_{max}])) \end{aligned} \tag{4}$$

4.1.2 Structural Reparameterization Module

In general, two effective ways to increase the proficiency of a network model are to improve its width and depth. Wider networks with more channels can extract more complex features. Deeper networks with more layers can learn more abstract and high-level representations of the data. However, increasing width and depth also increases computation and risk of overfitting, so these factors need to be balanced carefully when designing and training neural network architectures. The goal is to find the premier mixture of width and depth for the problem and dataset. Kaiming He et al. [42] in their research found that if we blindly add more layers to the neural network architecture, it will not only lead to an increase in parameters and a cumbersome model, but also lead to degradation issues, that is, network performance follows a curve - increasing depth helps up to an ideal level, after which it rapidly impairs results, and this phenomenon is not caused by overfitting, so He et al. made the network achieve constant mapping by adding residual branches, which allowed data to flow across layers. Thus He et al. proposed a series of new technologies, mainly through residual connections and batch normalization methods, to mitigate the degradation issue that may occur when simply stacking more layers to increase network depth. Literature [43] confirmed thatthe reason for the better performance of ResNet is that the branching structure of ResNet generates an implicit collection with a significant amount of submodels, which can enhance the model's learning capacity by establishing interdependent connections. These implicit sets can enhance the model's learning capacity through interdependencies, which all indicate that the multi-branch architecture has advantages that the single-way architecture model does not have.

However, in recent years, the networks are more oriented towards industrial deployment, and Christian et al. [44–47] proposed that for multi-branch networks, although the multi-branch structure can significantly improve the network accuracy, the operation speed of each layer depends on the branch with the slowest computation speed at each layer, which also leads to too many branches greatly slowing down the performance of the network and is not conducive to model deployment.

The emergence of structural reparameterization, an idea proposed by Ding et al. [48] at CVPR 2021, is a good solution to these problems. This indicates utilizing a multiple-branched architecture throughout the training procedure, and then combining the branches into a sole coalesced structure when performing inference, and this approach provides the benefits of both multi-branch and single-branch models. The multi-branch structure enables strong feature learning during training. The single-branch consolidation allows for fast and efficient inference after training. So it enables training a high-capacity model that learns rich features, while also obtaining fast prediction speed at test time.

Within this work, we introduce a novel module named RepConv-One Block that leverages the concept of multi-branched learning and consolidated deduction. We incorporate the RepConv-One Block into the backbone architecture of YOLOv5, which has a simpler structure and is easier to train compared with the existing reparameterization module. The specific implementation is to add a 1*1 residual branch to the backbone 3*3 size Conv module during training, and to restructure the model's parameters following the conclusion of training. The reparameterization process is presented in Figure 4, where in the batch normalization layer is first incorporated into the convolutional block, followed by converting the 1*1 convolution into a 3*3 convolution, and the last two fusing the 3*3 convolutions into one 3*3 convolution. The structure of the final obtained RepConv-One Block module during training is illustrated in Figure 5(a), while the multi-branch module is unified into a single convolutional stratum, as exhibited in Figure 5(b).

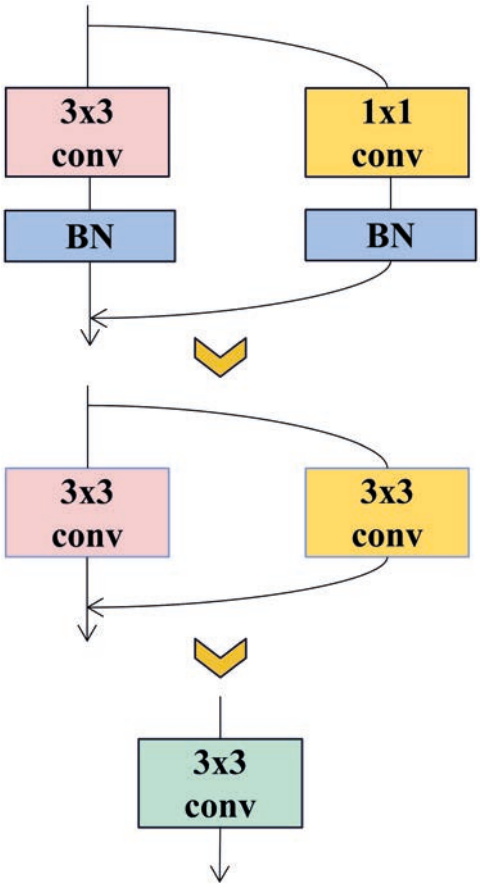

Figure 4. Reparameterization Process.

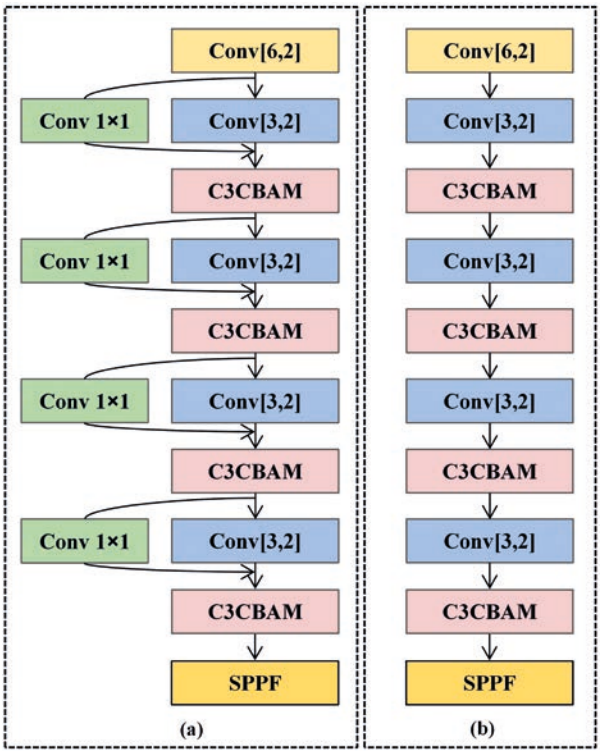

Figure 5. The Improved Backbone Structure.

The paper provides the specific derivations for fusing the multiple convolutional branches and batch normalization layers into a single consolidated convolutional block.

$$Conv(x) = W * x + b \tag{5}$$

$$BN(x) = \gamma * (x - Mean)/(\sqrt{\delta^2 + \varepsilon}) + \beta \tag{6}$$

where W represents the weight parameters, b denotes the bias term, in equation (6), β and γ represent learnable coefficients, $Mean$ is the mean across each batch of sample data, and σ is the variance, bringing the convolution layer to the BN layer yields equation (7)

$$BN(Conv(x)) = \gamma \times (W * x + b - Mean)/(\sqrt{\delta^2 + \varepsilon}) + \beta \tag{7}$$

or, equivalently

$$BN(Conv(x)) = \gamma * W * x/\sqrt{\delta^2 + \varepsilon} + \gamma * (b - Mean)/(\sqrt{\delta^2 + \varepsilon}) + \beta \tag{8}$$

After deriving the mathematical equations to consolidate the multiple convolutional branches and batch normalization layers, the authors arrive at equation (9) and (10) to represent the fused convolutional layer.

$$\begin{aligned} W_{\text{Fused}} &= \gamma * W/\sqrt{\delta^2 + \varepsilon}, \\ b_{\text{Fused}} &= \gamma * (b - Mean)/(\sqrt{\delta^2 + \varepsilon}) + \beta \end{aligned} \tag{9}$$

$$BN(Conv(x)) = W_{Fused} * x + b_{Fused} \tag{10}$$

Experiments show that the above strategy reduces the delay in hardware and significantly improves the accuracy of the algorithm without adding additional overhead.

4.2 Improvement of Upsampling Method

The current conventional upsampling methods can be summarized into three categories: linear interpolation-based upsampling method, deep learning-based upsampling method, and inverse pooling upsampling method. The underlying principles of these interpolation techniques are similar to manual feature engineering, As depicted in Figure 6, YOLOv5 specifically utilizes nearest neighbor interpolation for

upsampling feature maps in its backbone architecture. Although the nearest neighbor interpolation method is less overhead and faster, the feature map obtained by nearest neighbor interpolation method will break the asymptotic relationship of the original features are altered to some degree, which impacts the network's accuracy.

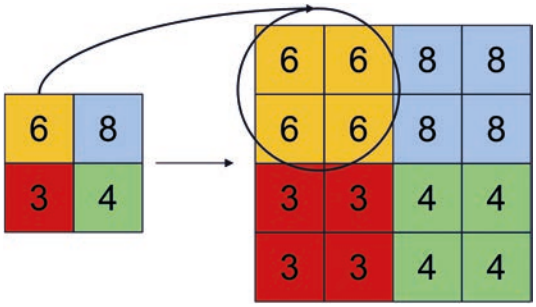

Figure 6. The upsampling method using nearest neighbor interpolation.

In this paper, we will use transposed convolution layers for upsampling instead of nearest neighbor interpolation. Specifically, we apply 4x4 inverted convolutions with a step size of 2 to enlarge the feature maps in the network, the principle of which is shown in Figure 7, equations (11) and (12) can represent the dimensions of the feature map after calculating the transposed convolution procedure. As opposed to the original interpolation methods, the transposed convolution upsampling has an additional layer of features that can be learned, enabling the network to more effectively learn high-level semantic representations of the image.Experiments demonstrate that using transposed convolutions for upsampling leads to significantly higher detection accuracy compared to nearest neighbor interpolation.

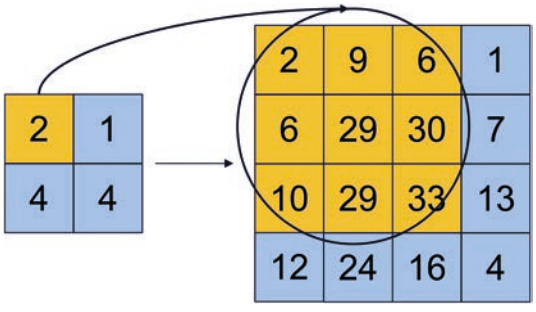

Figure 7. The upsampling method using transposed convolution.

$$H_{out} = (H_{in} - 1) \times stride[h] - 2 \times padding[h] + kernel\_size[h] \tag{11}$$

$$W_{out} = (W_{in} - 1) \times stride[w] - 2 \times padding[w] + kernel\_size[w] \tag{12}$$

4.3 Improvement of Detection Head

Song et al. [49] proposed that in today's popular object detection models, target classification and regression branches have different targets, leading to this spatial mismatch problem. The classification task focuses on feature similarity to predict class labels for each spatial location. On the other hand, the regression task aims to reduce the discrepancy between the projected bounding boxes and the actual bounding boxes. Therefore, if the same feature map is used to classify task and regression task, dislocation effect will be generated.

Based on this problem, Liu et al. [50] decouple the detection head in YOLOX, and achieve excellent detection effect, but it brings a huge parameters and calculation. In this paper, the detection head proposed by Liu et al. is further improved, the decoupled head designed by Liu et al. is lightweight processing, and a more concise lightweight decoupled head is proposed. As delineated in Figure 8, the left side (a) is the architecture of the decoupled head conceived by Liu et al. and the right side (b) is the structure of the lightweight decoupled head submitted within this work, the difference in accuracy between the two decoupled heads is presented in Table 1 Decoupled head comparison experiment. Compare with the decoupled head proposed by Liu et al., the parameters and the GFLOPs decreased by 25% and 35% respectively, while losing a little precision.

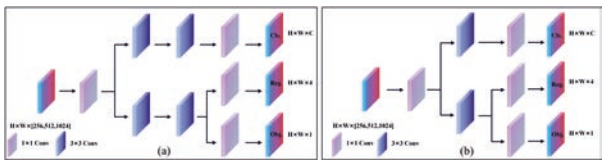

Figure 8. Comparison of Decoupled Head Structure.

Table 1. Decoupled head comparison experiment

Head	mAP0.5	mAP0.5:0.95	Params(10^6)	GFLOPs
Yolov5 Head	86.2	46.7	7.2	16.5
Decoupled Head	88.7(+2.5)	48.3(+1.6)	14.4	56.5
Concise Decoupled Head	88.1(+1.9)	47.8(+1.1)	10.8	36.7

4.4 Improvement of Loss Function

The YOLOv5 model optimizes three loss functions during training / a category mistake, a bounding box deviation, and a certainty discrepancy. More precisely, YOLOv5 employs Binary Cross Entropy (BCE) to calculate the classification and confidence losses. For the regression loss, Complete Intersection over Union (CIoU) is the loss function utilized by YOLOv5. This equation derives the CIoU loss from the mismatch between the prophesied bounding margin and the determinative ground truth margin:

$$L_{(CIoU)} = 1 - IoU + \rho^2(b, b^{gt})/c^2 + \alpha\upsilon \quad (13)$$

$$\alpha = \upsilon/((1 - IoU) + \upsilon) \quad (14)$$

$$\upsilon = 4/\pi^2 \times \left(\left(\arctan(w^{gt}/h^{gt})\right) - (\arctan(w/h))\right)^2 \quad (15)$$

In equation (13) $\rho(,)$ represents the Euclidean distance between the center points of the predicted bounding box (Bpred) and the ground truth bounding box (Bgt), C represents the length of the minutest diagonal box fully enclosing both Bpred and Bgt, b and b^{gt} represent the center coordinates of the Bpred and Bgt respectively, in equation (14) α denotes the weight function, υ represents a penalty term related to the image dimensions ratio similarity between the Bpred and the Bgt, w^{gt}/h^{gt} and w/h in equation (15) represent the dimensions ratio of the ground truth bounding boxes.

While the CIoU loss encapsulates multiple terms including intersection over union, centroid distance, and dimensions ratio consistency between the predicted bounding boxes and ground truth bounding boxes, though, the CIoU loss fails to factor in the rotational alignment of the forecasted box versus the authoritative box. This incapacity to penalize incongruity in box orientations can result in tardier convergence while training the model.

As outlined in this text, we utilize a loss function SIoU proposed by Zhora Gevorgyan [51] in 2022, as shown in Figure 9, SIoU considers the vector angle between the predicted bounding boxes and ground truth bounding boxes into the loss calculation, which is calculated as shown in equations (16) – (18). Where Δ symbolizes the distance cost and Ω symbolizes the shape cost for comparing the predicted bounding boxes and ground truth bounding boxes. The ablation experiments exhibit that SIoU significantly improves this model accuracy compared to CIoU.

$$L_{\mathrm{SIoU}} = 1 - IoU + (\Delta + \Omega)/2 \quad (16)$$

$$\Delta = \sum_{t=x,y}(1 - e^{-\gamma\rho_t}) \quad (17)$$

$$\Omega = \sum_{t=w,h}(1 - e^{-\gamma\rho_t})^{\theta} \quad (18)$$

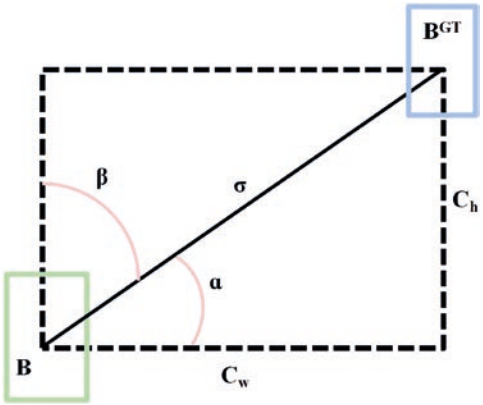

Figure 9. Scheme for calculating the angular cost in the loss function.

5 Experiments and Results

5.1 Experimental Environment

The experimental setup incorporates Ubuntu 20.04 as the underlying OS and Pytorch as the standard deep learning software framework, CUDA version 11.3.0, NVIDIA GeForce RTX 3090 as the GPU, 24 GB of video memory, 64 as the training batch size, and 300 epochs

in total. Before training, adjust the size of the input image to 640 * 640 pixels. Ground truth boxes were clustered using k-means to generate anchors. Data augmentation was performed by applying 4-Mosaic to expand the quantity of training data. Model optimization was performed using stochastic gradient descent (SGD) with a learning rate schedule that linearly increased initially before decaying.

5.2 Dataset Introduction

To validate our algorithm, we use the KAIST dataset at nighttime: KAIST [52]. The KAIST dataset comprises a total of 95,328 images, with each image containing both the RGB and infrared versions, resulting in a total of 103,128 densely annotated instances. Most papers utilizing the KAIST dataset have undergone preprocessing, as the dataset is derived from consecutive video frames with minimal differences between adjacent frames, hence necessitating a certain level of cleaning. This study focuses primarily on pedestrian detection in low-light environments, and as such, only the night dataset of KAIST was selected for experimentation.

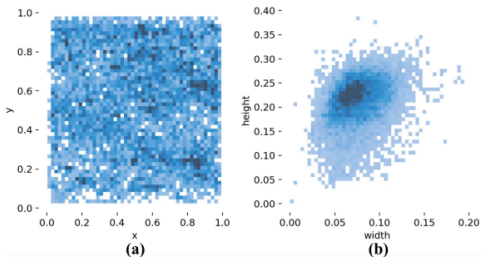

Figure 10. The Target Centroid Distribution.

Figure 11. The Target Size Distribution.

The cleaning process involved the following rules: (1) For the training set, every 2 images were selected, every 3rd image was chosen, and all images containing no pedestrians were removed, ensuring that the selected images each contain at least one target while excluding heavily occluded images showing only partial bodies or less than 50 pixels. (2) For the testing set, every 19 images were selected, i.e., every 20th image was chosen, and negative sample images were retained. Following the aforementioned operations, 5846 training set images and 1797 testing set images were obtained. The target centroids and target size distribution of the dataset are exhibited in Figure 10 and Figure 11.

5.3 Evaluation Indicators

Model assessment was based on mAP@0.5, mAP@0.5:0.95, and frames per second (FPS), number of parameters (Params), number of calculations (GFLOP), and model size (size) detected by the model. Among them, mAP0.5 represents the average AP at the IoU threshold of 0.5, and mAP0.5:0.95 represents the mean average precision obtained at IoU thresholds between 0.5 and 0.95 at 0.05 intervals, the calculation of which is given by the equations.(19–22).

$$\text{Precision} = TP/(TP+FP) = TP/\text{all detections} \tag{19}$$

$$\text{Recall} = TP/(TP+FN) = TP/\text{all ground truths} \tag{20}$$

$$AP = \int_0^1 P\,dR \tag{21}$$

$$mAP = \sum_{i=1}^{N} AP_i/N \tag{22}$$

To further validate the effectiveness of the proposed improvements and justify the introduction of CBAM, an extended set of ablation experiments was conducted. A total of 14 ablation models were designed to evaluate five core modules—RepConv, CBAM, transposed convolutional upsampling, Concise-Decoupled Head (CDH), and SIoU loss—and to compare CBAM with alternative attention mechanisms, including SE, ECA, and lightweight self-attention.

All models were trained under identical settings for fairness. The results are summarized in Table 2. As observed, each mod-

ule contributes differently to model performance. Notably, the inclusion of CBAM yields a substantial improvement in detection accuracy (mAP@0.5 increased by 0.9% compared to SE and 0.7% compared to ECA) while maintaining nearly identical computational cost. This advantage arises from CBAM's sequential channel-spatial attention design, which refines both feature discriminability and positional awareness under complex lighting conditions.

From Table 2, it can be concluded that CBAM provides the most significant accuracy enhancement among all tested attention modules, demonstrating its effectiveness in improving the robustness of feature extraction under complex night-time conditions without introducing excessive computational cost. Therefore, CBAM was retained as the optimal attention mechanism in the proposed CDH-YOLO framework.

5.4 Comparison Experiment

To provide supplementary evidence for the considerable merits of the presented CDH-YOLO algorithm, popular network models were utilized as comparisons in this study, such as YOLOv3 (2018), YOLOv3 tiny (2018), YOLOv5s (2020), YOLOv5 ghost (2020), YOLOv5m (2020), YOLOv6s (2022), YOLOv7 tiny (2022) and YOLOv8n (2023) on the KAIST dataset . We also compared with Transformer-based detectors (DETR) and traditional convolutional detectors (Faster R-CNN) to provide a comprehensive evaluation. And also compared with two popular models proposed in 2023: Improved-YOLOv7 tiny [53] and YOLO-Person [54]. The results of the experiment are presented in Table 3. The experimental results outlined above suggest the CDH-YOLO algorithm put forth in this manuscript achieves the highest 92.1% among the compared algorithms for the pedestrian detection task in night scenes mAP0.5, compared with YOLOv5s algorithm, CDH-YOLO architecture expands the parameter count by 57% and elevates computational intensity by 36% over the baseline YOLO model, mAP0.5 improves by 5.9%, mAP0.5:0.95% improved by 2.2%, where mAP0.5 and FPS have exceeded YOLOv5m by 2.5%, but CDH-YOLO only has half the number of parameters compared to YOLOv5m.

In addition, we built our proposed network on YOLOv8n, resulting in CDH-YOLOv8n, as shown in Table 3. Comparative experiments demonstrate that mAP0.5 and mAP0.5:0.95 are improved by 2.5% and 0.4%. Therefore, we can conclude that if the network is built on YOLOv8, the performance will still improve, further validating the robustness .

It is experimentally verified that CDH-YOLO achieves 64 FPS inference speed on a single GPU of NVIDIA GeForce RTX 3090, which satisfies the requirement of pedestrian detection task for real-time model. Combining the evaluation indexes of the number of parameters, computation volume, FPS, and mAP, CDH-YOLO achieves the best results and outmatches their capabilities.

5.5 Qualitative Evaluation

In this article, in order to contrast the detection capabilities of YOLOv5s and CDH-YOLO in different night scenes, three sets of images were used for the comparison purposes. The detection capabilities of YOLOv5s are illustrated in Figure 12.

As shown in the side-by-side comparison in Figure 12, across the foremost collection of images, YOLOv5s has two missed targets, one of which has low visibility and the other has a small amount of occlusion, while CDH-YOLO accurately detects all the targets in this image, which proves that CDH-YOLO has stronger feature extraction ability, the second group of targets is more severely occluded by tree branches, and YOLOv5s still has missed detection. In the third group of images, the target visibility is low and the recognition is difficult, but CDH-YOLO still performs well, according to the analyses of the former experiments, the evidence leads to the conclusionthat CDH-YOLO has richer feature extraction ability and can effectively cope with the pedestrian detection at night.

Table 2. Extended ablation experiments including comparisons among different attention mechanisms.

Model	Rep	Attention	Up Sam.	CDH	SIoU	mAP @0.5	mAP @0.5:0.95	GFLOPs	Size (MB)	Params ($\times 10^6$)
YOLOv5s	—	—	—	—	—	86.2	46.7	16.5	13.7	7.2
Model A	✓	—	—	—	—	87.6	48.3	16.5	13.7	7.4
Model B1	—	SE	—	—	—	86.9	46.9	16.6	13.8	7.3
Model B2	—	ECA	—	—	—	87.1	47.0	16.6	13.8	7.3
Model B3	—	Self-Att.	—	—	—	87.2	47.1	16.9	13.9	7.4
Model B4	—	CBAM	—	—	—	**87.8**	**47.3**	16.6	13.8	7.3
Model C	—	—	✓	—	—	88.7	47.2	16.6	13.7	7.2
Model D	—	—	—	✓	—	88.1	47.8	36.7	20.9	10.8
Model E	—	—	—	—	✓	87.7	46.8	16.5	13.7	7.2
Model F	✓	CBAM	—	—	—	88.2	48.2	17.7	14.8	7.6
Model G	✓	CBAM	✓	—	✓	89.1	48.4	18.7	15.1	7.8
Model H	✓	—	—	✓	—	89.3	48.5	36.9	21.3	11.0
Model I	✓	CBAM	—	✓	—	89.5	48.6	38.3	22.0	11.3
Model J	✓	CBAM	—	✓	✓	90.5	48.7	38.3	22.0	11.3
Model K	✓	CBAM	✓	✓	—	90.7	48.8	38.3	22.0	11.3
CDH-YOLO	✓	CBAM	✓	✓	✓	**92.1**	**48.9**	38.3	22.0	11.3

Table 3. Contrasting the detection performance across various algorithms

Models	Image size	Params(10^6)	mAP0.5	mAP0.5:0.95	FPS	GFLOPs
YOLOv5s (2020)	640×640	7.2	86.2	46.7	79	16.5
YOLOv3 (2018)	640×640	61.5	82.9	43.0	47	154.9
YOLOv3 tiny (2018)	640×640	8.6	77.9	34.6	179	12.9
YOLOv5 ghost (2020)	640×640	3.6	84.9	42.9	75	8.1
YOLOv5m (2020)	640×640	20.9	88.9	50.5	51.3	47.9
YOLOv6s (2022)	640×640	17.2	89.3	48.5	61	85
YOLOv7 tiny (2022)	640×640	6.2	88.1	46.5	286	13.8
Improved-YOLOv7 tiny(2023)	640×640	9.1	91.2	50.1	-	-
YOLO-Person(2023)	640×640	-	91.7	-	-	39
YOLOv8n (2023)	640×640	3.0	87.6	47.5	263	8.2
DETR (2020)	640×640	41.2	85.3	45.1	28	86.5
Faster R-CNN (2015)	640×640	137.5	83.7	44.3	15	203.7
CDH-YOLOv8n	640×640	8.3	90.1	47.9	217	19.1
CDH-YOLO	640×640	11.3	92.1	48.9	64	38.3

Figure 12. Side-by-side comparison of YOLOv5s Detection Effect.

6 Conclusion

This paper introduces CDH-YOLO, an enhanced YOLO-based framework designed for real-time pedestrian detection under nighttime conditions. The proposed framework incorporates structural reparameterization to optimize the backbone, CBAM attention mechanisms to strengthen feature representation, transposed convolution for semantically consistent upsampling, a lightweight decoupled head to mitigate spatial misalignment, and the SIoU loss function to achieve faster convergence and more precise localization. Extensive experiments conducted on the KAIST dataset demonstrate that CDH-YOLO achieves state-of-the-art accuracy while maintaining real-time inference speed, significantly surpassing existing approaches.

Although a significant accuracy improvement is obtained, this algorithm increases the amount and intricacy of calculations required by the model.The next stage of work will focus on two main aspects: Firstly, reducing unnecessary parameters and connections in the model through pruning, then transferring the knowledge from a large model to a smaller one through distillation. Through these methods, we aim to better explore lightweight processing of the model, improving its efficiency and performance to meet the increasing demands for pedestrian detection in night-time scenarios.

Conflict of 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.

Author contributions

Yanhua Ma: Conceptualization, Methodology, Writing-review and editing. Ke Lv: Software, Formal analysis. Li-Juan Liu: Writing-original draft, Data curation, Investigation. Hamid Reza Karimi: Visualization, Validation, Writing-review and editing.

Ethical and informed consent for data used

This research does not contain any studies with human participants or animals performed by any authors.

Data availability and access

The datasets analysed during the current study are available in the public resources: https://soonminhwang.github.io/rgbt-ped-detection/

References

[1] K. Min, G.-H. Lee, S.-W. Lee, Attentional feature pyramid network for small object detection, Neural Networks, 155, 2022, 439–450.

[2] C. Zhang, Q. Gao, R. Shi, M. Yue, LDHD-Net: a lightweight network with double branch head for feature enhancement of UAV targets in complex scenes, International Journal of Intelligent Systems, 2024, 7259029.

[3] J. Wei, S. Su, Z. Zhao, et al., Infrared pedestrian detection using improved UNet and YOLO through sharing visible light domain information, Measurement, 221, 2023, 113442.

[4] H. Shang, L. Sun, W. Qin, Pedestrian detection at night based on infrared camera and millimeter wave radar fusion, Journal of Sensor Technology, 34, 2021, 1137–1145.

[5] Y. Xue, Z. Ju, Y. Li, W. Zhang, MAF-YOLO: multi-modal attention fusion based YOLO for pedestrian detection, Infrared Physics & Technology, 118, 2021, 103906.

[6] L.-J. Liu, Y. Zhang, H. R. Karimi, Resilient machine learning for steel surface defect detection based on lightweight convolution, International Journal of Advanced Manufacturing Technology, 134, 2024, 4639–4650.

[7] L.-J. Liu, S.-Q. Sun, H.R. Karimi, A real-time surface defect detection model based on adaptive feature information selection and fusion, Information Fusion, 129, 2026.

[8] L. Zhang, L. Zhong, et al., Knowledge-guided multi-task attention network for survival risk prediction using multi-center computed tomography images, Neural Networks, 152, 2022, 394–406.

[9] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 42, 2017, 2011–2023.

[10] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: convolutional block attention module, Proceedings of the European Conference on Computer Vision (ECCV), 2018, 3–19.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, European Conference on Computer Vision, 2020, 213–229.

[12] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, GCNet: non-local networks meet squeeze-excitation networks and beyond, 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, 1971–1980.

[13] X. Ma, K. Hu, X. Sun, S. Chen, Adaptive attention module for image recognition systems in autonomous driving, International Journal of Intelligent Systems, 2024, 3934270.

[14] L.-J. Liu, Y. Zhang, H.R.Karimi, Defect detection of printed circuit board surface based on an improved YOLOv8 with FasterNet backbone algorithms. SIViP 19, 89 (2025).

[15] L.-J. Liu, S.-Q. Sun, H.R. Karimi, A real-time surface defect detection model based on adaptive feature information selection and fusion, Information Fusion, 129, 2026.

[16] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, OverFeat: integrated recognition, localization and detection using convolutional networks, CoRR, 2013.

[17] W. Liu, D. Anguelov, D. Erhan, et al., SSD: single shot MultiBox detector, European Conference on Computer Vision, 2016.

[18] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You only look once: unified, real-time object detection, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 779–788.

[19] R. B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2013, 580–587.

[20] R. Girshick, Fast R-CNN, 2015 IEEE International Conference on Computer Vision (ICCV), 2015, 1440–1448.

[21] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

[22] K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask R-CNN, IEEE Transactions on Pattern Analysis and Machine Intelligence, 42, 2017, 386–397.

[23] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 2015, 1137–1149.

[24] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 6517–6525.

[25] J. Redmon, YOLOv3: an incremental improvement, ArXiv, 2018.

[26] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, YOLOv4: optimal speed and accuracy of object detection, ArXiv, 2020.

[27] G. Jocher, YOLOv5 by Ultralytics, Available at https://github.com/ultralytics/yolov5, 2022.

[28] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., YOLOv6: a single-stage object detection framework for industrial applications, arXiv preprint arXiv:2209.02976, 2022.

[29] 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, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 7464–7475.

[30] W. Zhang, Z. Hong, L. Xiong, Z. Zeng, Z. Cai, K. Tan, Sinextnet: a new small object detection model for aerial images based on PP-YOLOE, Journal of Artificial Intelligence and Soft Computing Research, 14, 2024.

[31] X. Ji, J. Chang, Y. Ji, Adaptive separation fusion: a novel downsampling approach in CNNs, Journal of Artificial Intelligence and Soft Computing Research, 15, 2025.

[32] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, S. Belongie, et al., Feature pyramid networks for object detection, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 936–944.

[33] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 8759–8768.

[34] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, CoRR, 2014.

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. S. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, CoRR, 2015.

[36] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, et al., DRAW: a recurrent neural network for image generation, ArXiv, 2015.

[37] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, Advances in Neural Information Processing Systems, 28, 2015.

[38] J. Xu, Y. Cai, X. Wu, X. Lei, Q. Huang, H. F. Leung, Q. Li, Incorporating context-relevant concepts into convolutional neural networks for short text classification, Neurocomputing, 386, 2020, 42–53.

[39] Y. Cai, Q. Huang, Z. Lin, J. Xu, et al., Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: a multi-task learning approach, Knowledge-Based Systems, 203, 2020, 1–12.

[40] T. Hussain, W.-C. Wang, M. Gogate, K. Dashtipour, Y. Tsao, X. Lu, A. Ahsan, A. Hussain, A novel temporal attentive-pooling based convolutional recurrent architecture for acoustic signal enhancement, IEEE Transactions on Artificial Intelligence, 3, 2022, 833–842.

[41] M. Nawaz, T. Nazir, A. Javed, M. F. Masood, J. Rashid, J. Kim, A. Hussain, A robust deep learning approach for tomato plant leaf disease localization and classification, Scientific Reports, 12, 2022.

[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 770–778.

[43] A. Veit, M. J. Wilber, S. Belongie, Residual networks behave like ensembles of relatively shallow networks, Advances in Neural Information Processing Systems, 29, 2016.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, et al., Going deeper with convolutions, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, 1–9.

[45] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, ArXiv, 2015.

[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, et al., Rethinking the inception architecture for computer vision, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 2818–2826.

[47] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, Inception-ResNet and the impact of residual connections on learning, ArXiv, 2016.

[48] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, RepVGG: making VGG-style ConvNets great again, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 13728–13737.

[49] G. Song, Y. Liu, X. Wang, Revisiting the sibling head in object detector, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 11563–11572.

[50] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: exceeding YOLO series in 2021, ArXiv, 2021.

[51] Z. Gevorgyan, SIoU loss: more powerful learning for bounding box regression, ArXiv, 2022.

[52] S. Hwang, J. Park, N. Kim, Y. Choi, I. S. Kweon, Multispectral pedestrian detection: benchmark dataset and baseline, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 1037–1045.

[53] Y. Cao, C. Li, Y. Peng, Night pedestrian detection algorithm based on improved YOLOv7, Changjiang Information & Communications, 35, 2023, 57–60.

[54] Z. He, G. Chen, J. Chen, Y. Zhang, et al., Multi-scale feature fusion lightweight real-time infrared pedestrian detection at night, Chinese Journal of Lasers, 49, 2023, 1709002.

Yan-hua Ma received her PhD in Control Theory and Control Engineering from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2011. From 2012 to 2016, she was a deputy researcher with the Key Laboratory of Chemical Laser, Chinese Academy of Sciences. She is currently a professor with the School of Integrated Circuits, Dalian University of Technology, China. Her research interests include computer vision, software hardware co-design for neural network acceleration, fault diagnosis and fault-tolerant control of aero-engines, and control system design and application.
https://orcid.org/0000000222543896

Ke Lv received the B.S. degree in Software Engineering from Anhui Information Engineering University, Wuhu, China, in 2018. He is currently working toward the M.S. degree in Software Engineering with the School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian, China, and is expected to graduate in 2026. His research interests include computer vision, YOLO architectures, and object detection.
https://orcid.org/0009-0007-2354-4050

Li-Juan Liu received the M.S. degree in computer engineering from Liaoning Normal University, Dalian, China, in 2004, and the Ph.D. degree in control theory and control engineering with the Dalian University of Technology in 2019. She is currently a professor in the school of Information Science, Beijing Language and Culture University, China. Her current research interests include deep learning, machine learning, switching systems, time delay systems, positive systems, multi-agent systems and distributed control.
https://orcid.org/0000-0002-5716-2129

Hamid Reza Karimi is a Professor of Applied Mechanics in the Department of Mechanical Engineering at Politecnico di Milano, Milan, Italy. Prof. Karimi's original research and development achievements span a broad spectrum within the topic of automation/control systems, and intelligence systems with applications to complex systems such as vehicles, robotics and mechatronics. Prof. Karimi is an ordinary Member of Academia Europaea (MAE), Member of European Academy of Sciences and Arts (EASA), Member of The European Academy of Sciences (MEurASc), Member of Agder Academy of Science and Letters in Norway, Member of National Academy of Artificial Intelligence (NAAI), Honorary Academic Member of National Academy of Sciences of Bolivia, Distinguished Fellow of the International Institute of Acoustics and Vibration (IIAV), Fellow of The International Society for Condition Monitoring (ISCM), Fellow of the Asia-Pacific Artificial Intelligence Association (AAIA), and also a member of the IFAC Technical Committee on Mechatronic Systems, the IFAC Technical Committee on Robust Control, the IFAC Technical Committee on Automotive Control, member of the board of Directors of The International Institute of Acoustics and Vibration (IIAV) and member of Management Committee of The International Society for Condition Monitoring (ISCM). Prof. Karimi is the recipient of the 2025 NAAI Distinguished Artificial Intelligence Scholar Award, the 2021 BINDT CM Innovation Award, the Web of Science Highly Cited Researcher in Engineering, August-Wilhelm-Scheer Visiting Professorship Award, JSPS (Japan Society for the Promotion of Science) Research Award, and Alexander-von-Humboldt-Stiftung research Award, for instance. Prof. Karimi serves/served as Editor-in-Chief and Book Series Editor for Springer, CRC Press and Elsevier. He has also participated as General Chair, keynote/plenary speaker, distinguished speaker or program chair for several international conferences in the areas of Control Systems, Robotics and Mechatronics. Prof. Karimi has served as vice-president for International Prize "Lombardia è ricerca" and is also the conference chair for the 2026 World Congress on Condition Monitoring (WCCM2026) to be held in Milan, Italy. Additionally, he is an Honorary Visiting Professor in the School of Computing & Engineering at the University of Huddersfield, UK.
https://orcid.org/0000-0001-7629-3266