

Article

Research on Rapid Recognition of Moving Small Targets by Robotic Arms Based on Attention Mechanisms

Boyuan Cao, Aishan Jiang, Jiacheng Shen and Jun Liu *

School of Engineering and Technology, Southwest University, Chongqing 400715, China

* Correspondence: lcytsj@swu.edu.cn

Abstract: For small target objects on fast-moving conveyor belts, traditional vision detection algorithms equipped with conventional robotic arms struggle to capture the long and short-range pixel dependencies crucial for accurate detection. This leads to high miss rates and low precision. In this study, we integrate the traditional EMA (efficient multi-scale attention) algorithm with the c2f (channel-to-pixel) module from the original YOLOv8, alongside a Faster-Net module designed based on partial convolution concepts. This fusion results in the Faster-EMA-Net module, which greatly enhances the ability of the algorithm and robotic technologies to extract pixel dependencies for small targets, and improves perception of dynamic small target objects. Furthermore, by incorporating a small target semantic information enhancement layer into the multiscale feature fusion network, we aim to extract more expressive features for small targets, thereby boosting detection accuracy. We also address issues with training time and subpar performance on small targets in the original YOLOv8 algorithm by improving the loss function. Through experiments, we demonstrate that our attention-based visual detection algorithm effectively enhances accuracy and recall rates for fast-moving small targets, meeting the demands of real industrial scenarios. Our approach to target detection using industrial robotic arms is both practical and cutting-edge.

Keywords: attention mechanism; industrial robotic arm; small target; recognition; target detection; robot



Citation: Cao, B.; Jiang, A.; Shen, J.; Liu, J. Research on Rapid Recognition of Moving Small Targets by Robotic Arms Based on Attention Mechanisms. *Appl. Sci.* **2024**, *14*, 3975. <https://doi.org/10.3390/app14103975>

Academic Editor: Nikolaos Papakostas

Received: 14 March 2024

Revised: 28 April 2024

Accepted: 29 April 2024

Published: 7 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the rapid development of artificial intelligence technology, the use of robotic arms in fields such as industrial automation, healthcare, and aerospace is becoming increasingly widespread [1]. To achieve precise detection and tracking of targets by robotic arms, target tracking algorithms based on machine vision have emerged as a research hotspot. The attention mechanism originated from the study of human visual systems, allowing for the swift location of, and focus on, key information within complex environments [2,3]. Within the sphere of target tracking by robotic arms, the attention mechanism can assist in the rapid identification of targets and locking onto them within dynamic, complex environments. This results in improved accuracy and stability of tracking [4,5].

The active exploration of detecting and tracking small moving targets through machine vision represents a prominent research area within computer science [6]. Recent years have seen sustained interest and contributions from researchers worldwide, leading to continuous innovation and enhancement of algorithms for small moving target detection [7]. Given the intricate fluctuations present in dynamic detection environments—including variations in lighting and background noise—achieving high accuracy and robustness in dynamic target detection algorithms continues to pose significant challenges. Currently, the main approaches for small moving target detection algorithms include background elimination, optical flow, and inter-frame difference methods. Many scholars and institutions worldwide have conducted research into these small moving target detection algorithms and have yielded varying insights.

Singh, R [8] developed an image classification system that quickly categorizes a large number of factory-produced images into defective or non-defective classifications using

computer vision technology. The system utilizes Canny edge detection and SIFT (scale-invariant feature transform) matching algorithms for image matching. To obtain correctly classified photos, these images were compiled to create the test dataset provided to the system. Many of the enterprise's products have been covered by this program. However, the system is sensitive to variations in lighting, suggesting less environmental robustness.

Han, R [9] presented YOLO-SG, a salience-guided (SG) deep learning model that improves small object detection by attending to detailed regions via a generated salience map. YOLO-SG performs two rounds of detection: coarse detection and salience-guided detection. In the first round of coarse detection, YOLO-SG detects objects using a deep convolutional detection model and proposes a salience map utilizing the context surrounding objects to guide the subsequent round of detection. In the second round, YOLO-SG extracts salient regions from the original input image based on the generated salience map and combines local detail with global context information to improve the object detection performance. Experimental results have demonstrated that YOLO-SG outperforms the state-of-the-art models, especially when detecting small objects.

Du, J [10] posited that in the field of maritime navigation safety, it is important to identify whether certain specific items appear in the cockpit to help determine if there exists a threat to navigation safety. These items are generally small in size and necessitate higher detection efficiency. To address this issue, this thesis proposed a ship-specific item detection method that improved on the YOLOv5s algorithm. By introducing the CBAM (convolutional block attention module) [11], the proposed method enhanced the network's feature extraction capabilities and improved detection ability and accuracy for small targets. Experimental results showed that after the introduction of the attention mechanism, YOLOv5s achieved an accuracy of 85.6%, a recall rate of 85.2%, and an average accuracy of 90.2% for specific ship items, effectively accomplishing its task of detecting specific small objects in ship cabins.

The study illustrates how incorporating an attention mechanism can notably enhance the efficacy of detecting small targets. However, the reliance on the YOLOv5 framework suggests potential areas for further refinement in detection outcomes. Concentrating on deep learning-based methods for object detection and semantic segmentation in images, this research particularly addresses the classification of swiftly moving small objects on conveyor belts. It introduces an innovative detection and segmentation algorithm that leverages an attention mechanism, employing a deep learning-based modified basic network model. This methodology is effective for both identifying and segmenting stationary targets and tracking dynamic ones. Finally, the developed models were tested on common datasets for these detection tasks, and the improved algorithms achieved good experimental results. The main innovative points of this paper are as follows:

- (1) Given the challenges associated with accurately detecting small moving objects on high-speed conveyor belts in industrial settings—a task that conventional detection algorithms often struggle with—this study introduces an enhanced object detection algorithm aimed at reducing instances of missed detections and false positives.
- (2) The algorithm presented herein is founded on the YOLOv8 architecture and incorporates an attention mechanism as a plugin. This addition is designed to more effectively capture essential feature information, thereby ensuring that the refined algorithm meets the real-time demands of detection tasks.
- (3) With regards to the dynamic target detection problem, this paper employs a modified loss function to optimize the issue of poor detection results for dynamic small targets in the original algorithm.

2. Related Work

2.1. Introduction to the YOLOv8 Algorithm

Target detection algorithms are categorized into one-stage and two-stage detection algorithms, with the primary distinction being the two-stage algorithm's clear separation of processes: initially extracting feature regions, and subsequently identifying these re-

gions [11]. Methods such as RCNN (region-CNN) [12] are used to extract multiple feature regions, followed by the deployment of classification algorithms such as SVM (support vector machine) or deep neural networks to determine if they match the target object. These tactics are usually time-consuming, rendering them less suitable for scenes requiring high real-time rates. One-stage target detection algorithms integrate the aforementioned processes, predicting the target region directly from the original image. Among these, the YOLO (You Only Look Once) algorithm is the most representative one-stage target detection model [13], offering major speed improvements over R-CNN type algorithms and thus providing feasibility for online real-time applications. Since 2016, this method has received considerable attention, leading to multiple improvements. Upgraded versions include YOLOv5 [14], YOLOv8 [15], and YOLOv9 [16]. The latest algorithm in the YOLO series has been updated to v9, but its fundamental framework remains rooted in the core principles of the YOLO series. YOLOv9 primarily focuses on the design of programmable gradient information (PGI) and generalized ELAN (GELAN), without altering the basic structure of YOLO. Therefore, after careful comparison, we have opted to continue using YOLOv8 as our foundational structure, given its relatively lightweight nature and proven effectiveness across various computer vision tasks and scenarios. YOLOv8 provides various scale models, including N/S/M/L/X, to accommodate different scene requirements. Table 1 shows the parameters of various scales of the YOLOv8 model.

Table 1. YOLOv8 model parameters at various scales.

STRUCTURE	YOLOv8N	YOLOv8S	YOLOv8M	YOLOv8L	YOLOv8X
Depth	0.33	0.33	0.67	1.00	1.00
width	0.5	0.50	0.75	1.00	1.25
Max_channels	1024	1024	768	512	512

YOLOv8, as one of the most widely used versions in practical applications, not only inherits the features of previous versions, offering good recognition accuracy and fast recognition speed, but also enhances the performance of the model itself, enabling broader application in real production. Figure 1 presents the main structure of YOLOv8. The detection effect of YOLOv8 demonstrates significant improvement compared to other mainstream object detection models [17]. By improving the accuracy of the detection algorithm, YOLOv8 enables more accurate target identification. Early object detection was completed in simple environments; in the task of recognizing fast-moving small targets, improvements to the model are needed to accomplish identification tasks in more complex environments [18].

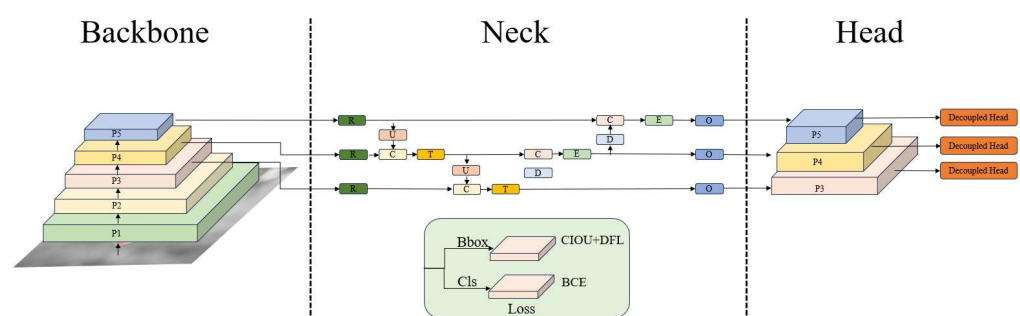


Figure 1. Structure of YOLOv8.

YOLOv8 replaces the C3 module in YOLOv5's backbone with the gradient-rich c2f module, reducing the number of blocks in the maximum stage of the backbone network. This further decreases the parameter amount and computational needs, realizing a more lightweight model [19]. YOLOv8 also eliminates the convolutional structure in the up-sampling stage of PAN-FPN from YOLOv5, thus speeding up computation. YOLOv8

employs an anchor-free detection approach, which predicts the center point and aspect ratio of the target directly instead of predicting the position and size of the anchor box. This method reduces the quantity of anchor boxes, leading to enhanced detection speed and accuracy [20].

This chapter gives a brief introduction to the new features of the new algorithm network.

1. We introduce and enhance a novel multi-scale parallel network structure, Faster-EMA-Net, which is based on the traditional EMA algorithm. It incorporates the c2f module from the original YOLOv8 and integrates a Faster-Net module designed based on partial convolution concepts. This module significantly reduces redundant computations and memory accesses, thus enabling more efficient acquisition of high-quality primary feature information. Consequently, it improves the accuracy of small object detection. The comparison diagram between the network structure of Faster EMA Net, the C3 module in YOLO v5, and the c2f module in YOLOv8 is shown in Figure 2.

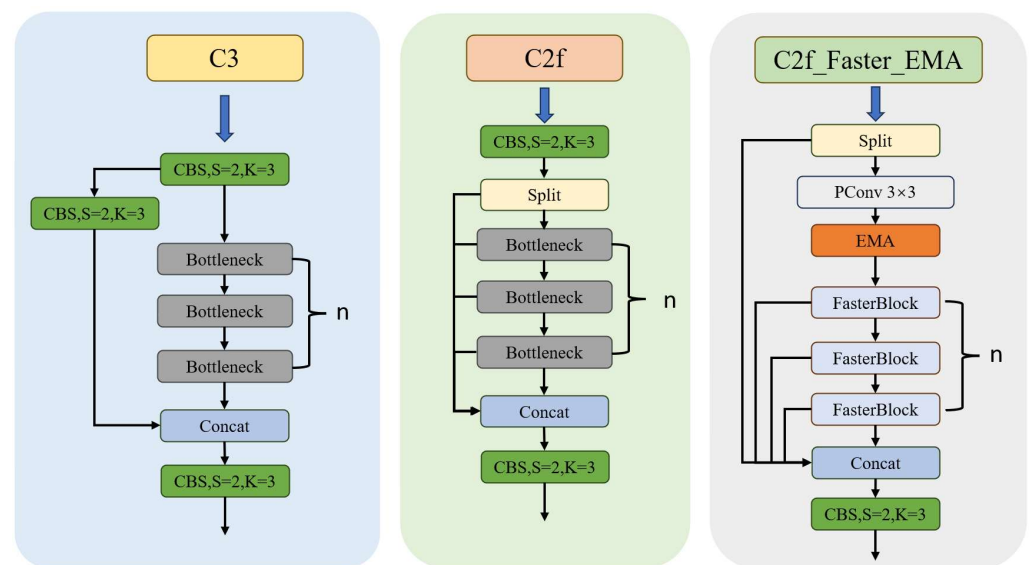


Figure 2. Comparison between C3 module and c2f module.

2. This paper improves the loss function of YOLOv8 to optimize the measurement of overlap between predicted boxes and ground truth boxes. This enhancement enables the model to better fit the object boundaries, thereby enhancing the accuracy and stability of dynamic object detection. The improved algorithm is better suited for real-time detection tasks.

3. To demonstrate the generality and effectiveness of our algorithm, detailed experiments are presented in Section 4. Three types of fruits on a conveyor belt are selected as experimental objects. Attention comparison experiments and ablation experiments on various parts of the improved network are conducted. The experimental results show that compared to the original network, our designed structure can effectively capture pixel dependencies in images, resulting in a significant improvement in the detection of small objects, while reducing both false positives and false negatives.

2.2. Introduction to the Attention Mechanism

In recent years, the attention mechanism has been widely used across various fields of deep learning.

Many scholars have applied attention mechanisms in the field of computer vision. Zhou et al. [21] proposed a novel end-to-end learning neural network called dynamic visual attention prediction (DVAP), which included an asymmetric attention block named motion attention transition (MAT) for zero-shot video object segmentation (ZVOS). They achieved promising results in practice. Yang et al. [22] introduced a multi-attention mechanism that addressed the issue of non-local attention blocks, including all features in the attention

matrix calculation. Despite having consistent input and output dimensions, this method can be seamlessly integrated into the structure of convolutional neural networks. Experimental results demonstrated that this multi-attention mechanism effectively improved the model's recognition accuracy, surpassing other neural network architectures.

Loo et al. [23] proposed an end-to-end depth feature fusion network for iris-periocular joint recognition, integrating various attention mechanisms including self-attention and cooperative attention into the network. They introduced a cooperative attention mechanism in the feature fusion module, which adaptively fused features to obtain more representative periocular features. The network with few parameters outperformed both unimodal biometric recognition and several iris-periocular recognition methods when applied to two publicly available datasets. Shi et al. [24] introduced Attention YOLOX, embedding attention mechanisms into the object detection algorithm YOLOX to improve detection accuracy. Evaluation on datasets confirmed that the proposed method achieved better detection accuracy than YOLOX without attention mechanisms, while maintaining high processing speed.

Dong et al. [25] addressed the challenges of dynamic object detection by proposing a method to enhance detection performance based on spatiotemporal attention mechanisms. They introduced the concept and advantages of spatiotemporal attention mechanisms and discussed the key techniques and fundamental issues of spatial attention mechanisms in dynamic object detection. They emphasized the potential effectiveness of dynamic object detection methods based on spatiotemporal attention mechanisms. Du et al. [26] proposed a multi-branch detection network model based on dual attention mechanisms, capable of detecting subtle artifacts. The algorithm's performance was evaluated on a deepfake dataset, with experimental results showing a testing accuracy of 96.45%. Moreover, compared to other methods, this approach also exhibited good performance in adversarial testing. Liang et al. [27] introduced a novel method integrating attention mechanisms into LBP and CNN networks to enhance facial expression recognition accuracy, validated on public datasets. The self-built attention mechanism helped the model automatically select and focus attention on the most informative regions of the image, thereby improving recognition performance on key areas and overall accuracy. Comparative experiments demonstrated that attention-based methods achieve higher recognition accuracy and robustness in facial expression recognition tasks, presenting clear advantages over traditional methods.

Wu et al. [28] proposed a Siamese network object tracking algorithm combined with attention mechanisms based on existing Siamese network algorithms. To improve model accuracy and tracking success rate, they introduced an attention mechanism module composed of channel attention and spatial attention mechanisms. Experimental results showed that the proposed Siamese network object tracking algorithm combined with attention mechanisms achieved favorable results on the OTB100 test dataset, outperforming the original method in terms of performance enhancement. Yang et al. [29] presented a model based on dual attention mechanisms. The self-attention mechanism effectively extracted dependencies between local and global features in images, while the channel attention mechanism automatically determined the importance of each feature channel, leading to more efficient resource allocation. Experimental results demonstrated that this method outperformed others on the CelebA dataset, producing higher-quality images. Chen et al. [30] proposed a novel ASPP structure with global attention information to provide distant detailed information more effectively for semantic segmentation models. Subsequently, a new SAM (selective attention module) was introduced in the decoder stage to provide different attention weight information for different spatial positions. Results demonstrated that both the ASPP with global attention and the decoder with a selective attention mechanism significantly improved accuracy.

The proposed attention mechanisms mainly exist in three types: channel attention, spatial attention, and those incorporating both [31]. The squeeze-and-excitation (SE) model serves as a representative of channel attention, explicitly modeling interdependencies across different channels to extract channel-wise attention. The convolutional block at-

tention module (CBAM) represents a simplistic yet effective feed-forward convolution neural network attention module [32]. However, CBAM can only extract local positional attention information, and lacks the ability to extract long-distance relationships. The coordinate attention (CA) mechanism is a new, efficient attention mechanism that takes both the channel and direction-related position information into account. The CA attention mechanism can provide significant benefits for downstream tasks based on lightweight networks. It allows attention blocks to capture long-distance relationships in one direction while retaining spatial information in the other. In this way, the positional information can be preserved in the generated attention maps which focus on areas of interest, helping the network to more effectively and accurately localize the targets [33]. The computational formula for the attention mechanism is as follows [34]:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

In Formula (1), Q represents Query, K represents Key, and V represents Value. Q , K , and V are linear transformations derived from the same input matrix X [35]. The attention function transforms each query into a weighted sum of values, where the weight is calculated by normalizing the dot product between queries and their corresponding keys. The fundamental computational formulas for Q , K , and V are as follows:

$$Q = XW^Q, K = XW^K, V = XW^V \quad (2)$$

In the above, W^Q , W^K , and W^V represent three trainable parameter matrices. The input matrix X , when multiplied with W^Q , W^K , and W^V , generates Q , K , and V , effectively undergoing a linear transformation [36]. The attention mechanism does not directly use X , but rather the three matrices generated through matrix multiplication, because utilizing three trainable parameter matrices can enhance the model's fitting capacity. Q and K^T undergo a matrix multiplication (MatMul) operation to yield a similarity matrix. Each element of the similarity matrix is then divided by $\sqrt{d_k}$, where d is the size of K 's dimension. This ensures a more stable gradient update during training. Following normalization via the Softmax function, each value is a weight coefficient greater than 0 and less than 1, with their sums equal to 0. This result can be understood as a weight matrix. Finally, the weight matrix just obtained is multiplied with V to calculate the weighted sum. The calculation process is illustrated in Figure 3.

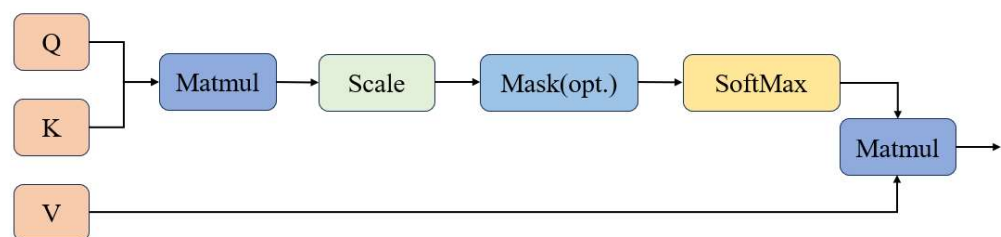


Figure 3. Process of attention mechanism calculation.

3. YOLOv8 Model Improvements

In the context of swift sorting tasks on conveyor belts, the belt typically moves at high speeds while a robotic arm remains stationary above the workstation. Cameras employ an “eye-outside-hand” technique to identify objects on the moving conveyor belt. Traditional algorithms, based on image processing are often applied to large, slow-moving targets in a stable illumination environment, but for fast-moving small targets, false detections and missed detections are often incurred [37]. This article improves the fast-moving small target recognition and grabbing algorithm based on two aspects: firstly, the EMA attention mechanism is incorporated within the network structure; secondly, the regression loss function is enhanced.

3.1. Incorporating Attention Mechanism in the Network Structure

Firstly, the problem of detecting fast-moving small targets can be defined as identifying the position and movement of one or more target objects in a series of continuous image frames [38]. In the context of robot arm applications, this usually involves the detection of specific objects to execute precise control and operations. The task is fraught with challenges due to the complexity of environments and the variability in target shapes, necessitating the incorporation of attention mechanisms. The attention mechanism mimics human visual focus, enabling the model to prioritize and process the most pertinent information for the task at hand while disregarding less relevant details. In the algorithm for moving small object detection, this implies that the algorithm can automatically recognize and focus on target objects in the image, thus improving detection accuracy and efficiency [39].

In this study, YOLOv8 is taken as the base model and improved to adapt to small target detection in dynamic environments. The improvements for YOLOv8 are as follows: by adding the EMA attention mechanism before the bottleneck in the c2f module of the backbone, the network can more accurately locate the target and improve network efficiency. The EMA structure is shown in Figure 4.

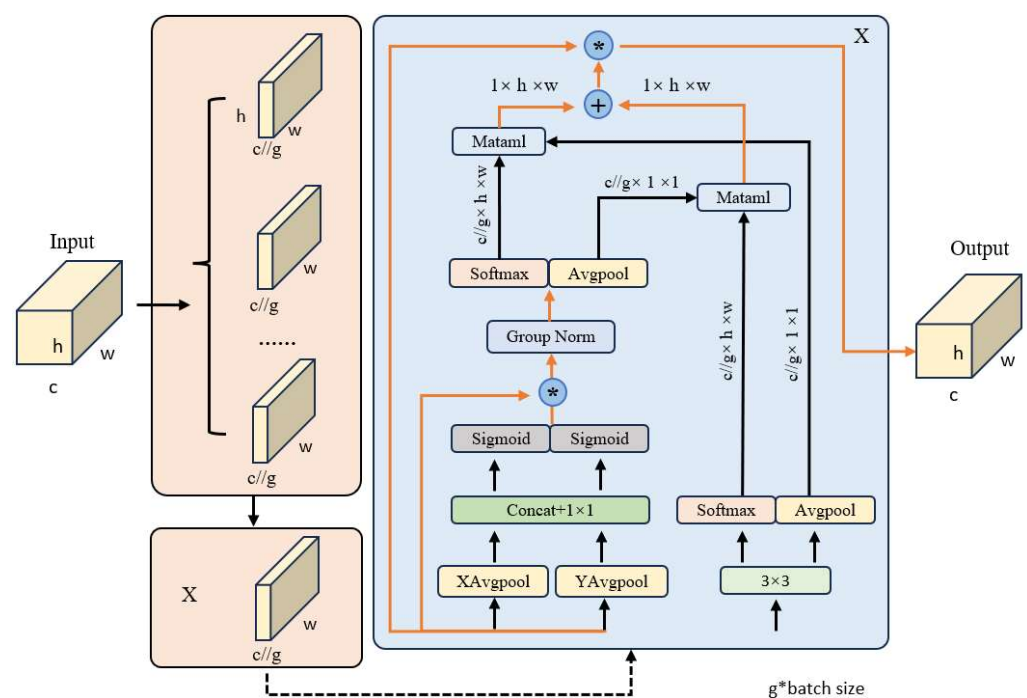


Figure 4. Structure diagram of the EMA attention mechanism.

This article proposes applying the EMA (efficient multi-scale attention) module to detect fast-moving small targets. EMA aims to preserve information in each channel and reduce computational overhead, partly reshaping the channels into batch dimensions, and grouping channel dimensions into multiple sub-features. This method enables spatial semantic features to be evenly distributed in each feature group.

This article integrates the EMA module with the improved c2f module. The modified version of the c2f module, based on fractional convolution design in Faster-Net, reduces redundant computations and memory access, thus more effectively extracting spatial features. By incorporating the EMA mechanism, the model enhances its sensitivity to targets that appear within the frame. This refinement to the c2f module is depicted in Figure 5. Among them, * represents dot product, + represents aggregation.

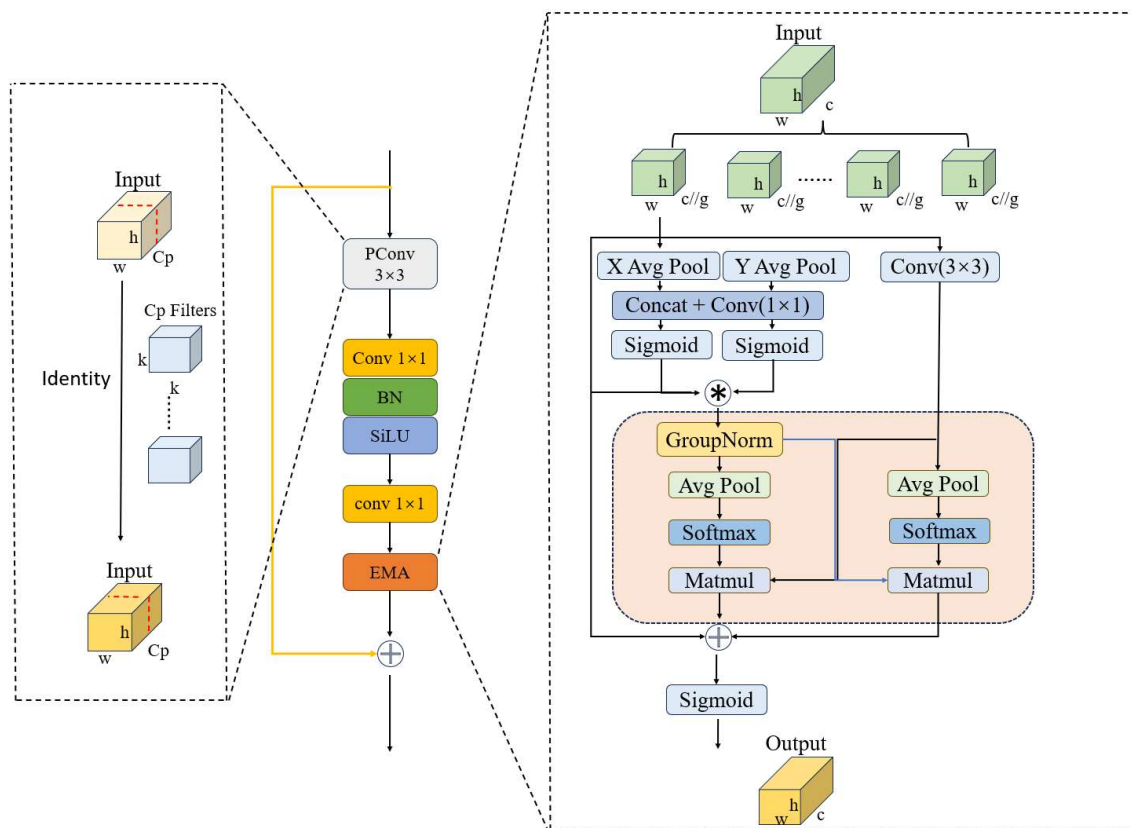


Figure 5. Improved c2f module.

3.2. Improving the Regression Loss Function

In the YOLOv8 model, the Complete IOU (CIOU) is used as the detection box regression loss function [40]. The model takes into account the overlapping area, center point distance, and aspect ratio between the predicted box and the actual box in bounding box regression. That is,

$$\{(W = kW_{gt}, H = kH_{gt}) | k \in \mathbb{R}^+\}$$
 (3)

In the equation, W and H denote the width and height of the predicted bounding box, respectively, while W_{gt} and H_{gt} stand for the width and height of the ground truth box. The aspect ratio within the formula simply indicates the proportional relationship between the width or height of the predicted box and that of the ground truth box. When the aspect ratios between various predicted boxes and actual boxes align, the computed result of the CIOU metric will be identical. Addressing this issue, in this study, the SIOU is adopted as the detection box regression loss function. Zhora Gevorgyan demonstrated that center-aligned bounding boxes converge more rapidly, forming the SIOU by composing the angle cost, distance cost, and shape cost. The ‘angle cost’ describes the smallest included angle between the line connecting the centers of the bounding boxes and the x - y axis:

$$\mathcal{A} = \sin \left(2 \sin \frac{\min(|x - x_{gt}|, |y - y_{gt}|)}{\sqrt{(x - x_{gt})^2 + (y - y_{gt})^2 + \mu}} \right)$$
 (4)

The term ‘distance cost’ describes the normalized distance between the center points of the two bounding boxes on both the x -axis and y -axis. Its penalty strength is directly proportional to the ‘angle cost’. ‘Distance cost’ is defined as follows:

$$\Delta = \frac{1}{2} \sum_{t=w,h} (1 - e^{-Y_{\rho t}}), Y = 2 - \mathcal{A}$$
 (5)

$$\begin{aligned}\rho_w &= \left(\frac{x - x_{gt}}{W_g} \right)^2 \\ \rho_h &= \left(\frac{y - y_{gt}}{W_g} \right)^2\end{aligned}\quad (6)$$

The term ‘shape cost’ describes the shape differences between the two bounding boxes. It is not equal to zero when the sizes of the two bounding boxes are inconsistent. ‘Shape cost’ is defined as follows:

$$\Omega = \frac{1}{2} \sum_{t=w,h} (1 - e^{wt})^\theta, \theta = 4 \quad (7)$$

$$\begin{aligned}\omega_w &= \left(\frac{|w - w_{gt}|}{\max(w, w_{gt})} \right) \\ \omega_h &= \left(\frac{|h - h_{gt}|}{\max(h, h_{gt})} \right)\end{aligned}\quad (8)$$

R_{SIOU} is similar to R_{CIOU} , as both are composed of a distance cost and a shape cost. Building upon the foundation of CIOU, SIOU introduces an additional consideration for shape similarity, which refines the loss calculation further. This enhancement allows for more effective handling of bounding box matches that exhibit significant shape differences. Thus, the focus is not only on matching position and size, but also on ensuring shape consistency. The formula for R_{SIOU} can be described as follows:

$$R_{SIOU} = \Delta + \Omega \quad (9)$$

For rapidly moving small targets, a grab algorithm is designed based on the improved YOLOv8 target detection method and depth ranging. This article uses the D435i camera to real-time capture both color images and depth images of grabbing scenarios. Through the target detection method applied to color images, the positional information of the target object is acquired, resulting in the output of the two-dimensional pixel coordinates (X, Y) for the grasping location. The depth value (Z) corresponding to this grasping position is derived from the depth image. These coordinates are then transformed into rotation angles relative to the various axes of the robot arm, guided by the coordinate transformation relationship. This process enables the robotic arm to precisely grasp the target.

3.3. Incorporating a Small Object Detection Layer into the Network Architecture

This section introduces the enhancement of the network structure with a dedicated layer for the detection of small objects, aiming to improve the system’s ability to identify and interact with smaller targets efficiently. According to relevant experimental results, the YOLOv8 model in itself generally exhibits average performance in detecting small targets. This is because these target samples are small in size, making their features more difficult to capture in relation to the overall image. YOLOv8 has high degrees and frequencies of down-sampling, making it harder for the deeper feature maps to learn the feature information of small targets. As we need to study the characteristics of even smaller objects, this article adds a small target detection layer, which uses the concatenation operation to merge shallow feature maps and deep feature maps for detection. The addition of a small target detection layer makes the network pay more attention to small object information, improving the actual results of small target detection.

The structure of the improved YOLOv8 model in this article is shown in Figure 6, with improvements marked in red.

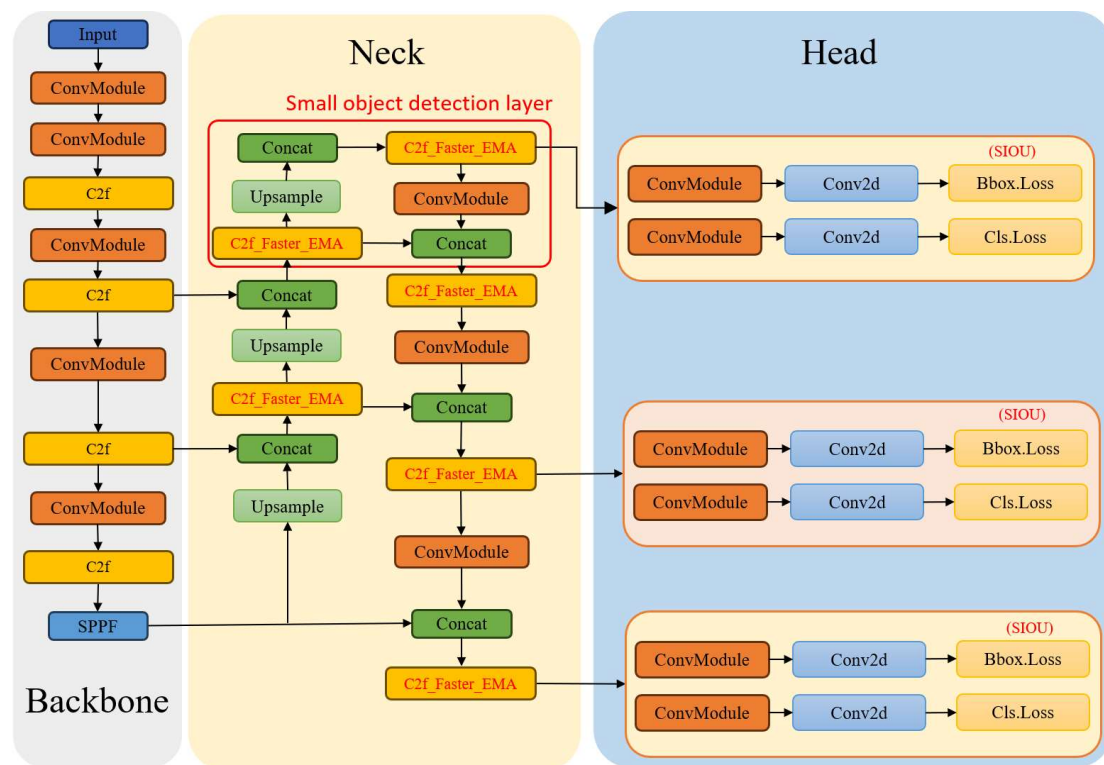


Figure 6. Structure diagram of the improved YOLOv8 network.

4. Experimental Results and Discussion

The improved algorithm was tested on the robotic arm's control system through practical experiments. The entire system was divided into two parts: hardware construction, and algorithm development with software implementation. In terms of hardware construction, the hardware used in our experiments is listed in Table 2.

Table 2. Configuration parameters.

Configuration Name	Version Parameters
Camera	RealSense-D435i
Robotic arm	Franka Emika Panda
GPU	GeForce RTX 4090 24GB
CPU	AMD EPYC 9654 96-Core Processor
operating system	Ubuntu 20.04
GPU acceleration library	Cuda11.0
programming language	Python3.8.10

The entire system was designed to identify three types of fruits—bananas, apples, and oranges—on a conveyor belt using the RealSense D435I depth camera. The process involved target detection through image data captured by the camera, ultimately pinpointing the central pixel coordinates (X, Y) of the small fruit targets, along with the depth information (Z) of point O. These coordinates were then transformed and transmitted to the Franka Emika Panda robotic arm. Upon receiving the instructions, the robotic arm moved to the location of the identified small target fruits and executed fruit sorting through its end effector. The actual experimental environment is shown in Figure 7.

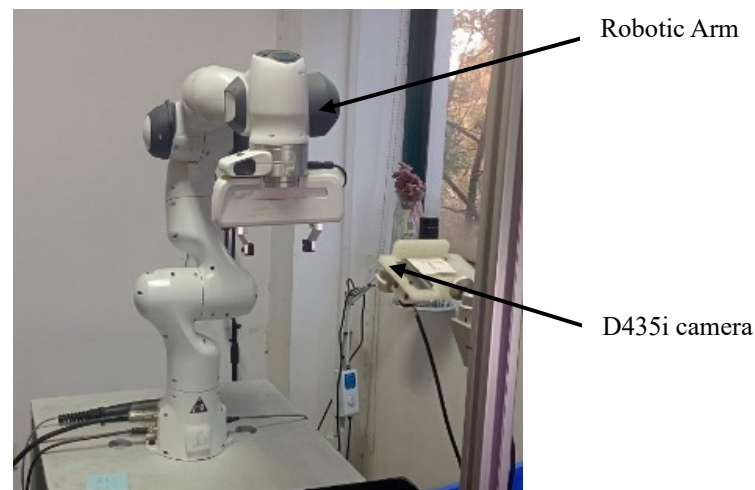


Figure 7. Experimental Scene Display.

In the execution of target detection and grasping tasks, real-time detection of the color image stream captured by the camera was required to meet the demands of real-time grasping. Once the vision-based grasping system was calibrated and the central pixel coordinates of the target objects were known, the position information in the base coordinate system was obtained through a series of coordinate transformations. Grasping operations were then carried out using inverse kinematics solutions and motion planning algorithms available in the ROS (robot operating system). This intelligent sorting system integrates various functional modules on the ROS platform. Based on ROS's topic subscription and publishing communication mechanism, the main functional modules of the system are designed as nodes, which communicate with each other by subscribing to or publishing relevant topics, ultimately achieving target grasping. Figure 8 shows an overall control diagram combining the robotic arm's hardware and software, with the control system's upper layer based on `Franka_ros`. Franka ROS integrates Franka Emika robots into the ROS system and incorporates Libfranka into ROS Control for controlling the robotic arm. The robotic arm's real-time transmission rate was 1kHz, satisfying the requirements for real-time grasping.

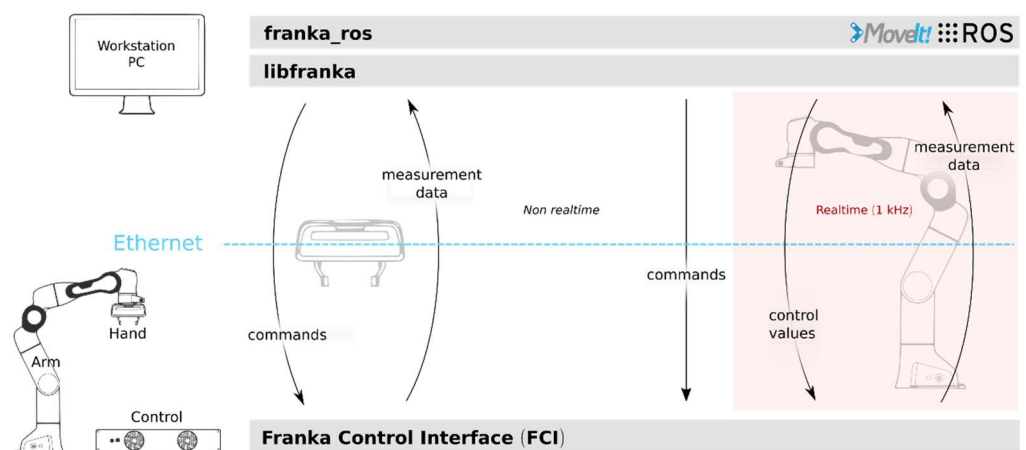


Figure 8. Overall diagram of robotic arm control.

To address these issues, we utilized built-in tools in ROS for monitoring and diagnostics. For example, we used ‘rostopic hz’ to monitor the publishing frequency of topics, ‘rosping’ to test the communication delay between nodes, ‘rqt_graph’ to view the relationship graph between nodes and topics, and ‘roswtf’ to detect issues within the system.

We ensured that all of the devices in the ROS network were on the same local network to reduce network delays. We used wired connections instead of wireless, because wireless can be unstable and have higher latency. During system operation, we closed or uninstalled unnecessary processes to reduce computer load. We optimized algorithms and data processing workflows to ensure code efficiency.

We adjusted the ‘ros::Rate’ object to control the running frequency of nodes and modified the message queue sizes to prevent the accumulation of too many messages in nodes.

Through these practices, our designed system underwent multiple iterations, adjustments, and tests to achieve optimal system performance. Ultimately, we successfully ensured the real-time effectiveness of our target detection and grasping processes.

4.1. Experimental Setup for Reliability

To ensure the reliability of the robotic arm operations, fruits were randomly placed at different heights and positions. Experimental results were referenced by the coordinates of the fruit targets. We conducted 100 sets of experiments under indoor lighting conditions, where the robotic arm performed identification and grasping tests. From these, we randomly selected 10 sets of fruit at different positions to gather data on detected coordinates and errors, as shown in Table 3.

Table 3. Group system reliability testing.

Group ID	Camera 3D Coordinate Values (X, Y, Z)/mm	Actual Measured Value Zr/mm	Error $\frac{ Z - Z_r }{Z_r} / \%$	Grab Results
1	(193.1, 66.2, 50.2)	49.5	1.55	Yes
2	(169.0, 82.6, 57.9)	57.2	1.32	Yes
3	(175.7, 67.6, 86.3)	83.6	3.26	No
4	(210.5, 78.5, 69.5)	67.9	2.36	Yes
5	(203.2, 53.2, 57.2)	56.3	1.63	Yes
6	(206.6, 77.9, 84.5)	82.1	3.01	Yes
7	(221.3, 86.2, 73.0)	71.0	2.85	Yes
8	(190.5, 59.6, 77.7)	76.2	2.06	Yes
9	(186.9, 65.4, 83.3)	60.9	1.98	Yes
10	(200.1, 86.8, 62.1)	59.6	1.88	Yes

The data from Table 1 indicate that each set of experiments showed varying degrees of error between the system-calculated values and the actual measured values. The largest error in a group was 3.26%, and the smallest was 1.32%. Due to the distance between the target and the camera, this might lead to failures in the robotic arm’s grasping attempts. However, overall, the experimental data showed a successful grasping rate of 96%. The overall errors met the precision requirements of the experiment and aligned with the expected objectives. This demonstrates that the system’s performance was robust and reliable for practical applications, even considering the potential inaccuracies due to positional variances and operational conditions.

The technical specifications of D435i are outlined in Table 4. From the table, it is evident that this camera features compact size, a wide field of view, high resolution, and easy installation. Equipped with Intel’s latest depth-sensing hardware and software, the camera boasts high integration. Intel’s official website offers the cross-platform development software Intel RealSense SDK 2.0, which provides a rich set of interfaces for secondary development.

Table 4. Technical specifications of D435i.

Parameter	Specifications
Scope of use/(m) interface	0.3~3.0 USB Type-C 3.1 Gen 1
Length, width, height/(mm * mm * mm)	90*25*25
Field of View/(°)	69.4*42.5
Resolution/Pixel	1920*1080

4.2. Dataset Preparation

This article used the RealSenseD435I depth camera to capture and search online to create a total of 5213 original photos of small target fruits, including apples, bananas, and oranges. Each photo contained a random number of experimental fruits.

The rotating boxes were annotated with the LabelImg tool, resulting in a custom dataset formatted similarly to COCO. This dataset was divided into training, validation, and testing subsets following a 7:2:1 ratio, respectively. To facilitate a comparative analysis, both the YOLOv8 standard model and the enhanced model proposed in this paper were trained using these specified training, validation, and test sets.

4.3. Experimental Environment Configuration

This paper used the PyTorch open-source deep learning framework for the development of the improved network model. After development was complete, model training and object detection experiments were conducted. The CPU used at the training end was the Intel Core i7-13700KF, and the GPU was the NVIDIA Geforce RTX3090. The batch_size was set to 16, base_lr was set at 0.008, α in RTAL was 1, and β was 6. The weights of the classification, SIOU, and DFL loss functions were set to 1.0, 2.5, and 0.05, respectively. CUDA11.2 and cudnn8.2 were used for acceleration during the training process. For the testing end, the same technology as at the deployment end was used for predicting small target fruits.

4.4. Comparative Analysis of Experimental Indicators

This study used precision-recall as one of the evaluation indicators. A high precision score suggests that most objects detected by the model are indeed objects, with only a few non-object entities being classified as objects. A high recall score, on the other hand, suggests that the model can locate more objects in the image. As can be seen, compared with the original network and YOLOv8-SE, both the precision and recall of the improved YOLOv8n were significantly higher than the other two network models.

Intuitively speaking, measuring the quality of object detection using precision-recall seems sufficient. However, in object detection, each image may contain different targets of different categories, implying that the model's classifications and localizations need different indicators for evaluation. *mAP* (mean average precision), commonly used in object detection to measure identification accuracy, was used as the reference indicator, representing the area underneath the precision–recall (P–R) curve. The formulas for *mAP*, *P*, and *R* are as follows:

$$mAP = \frac{\sum_{i=1}^k AP_i}{k} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

For variables in formulas, *k* represents the number of class objects. *TP* (true positives), *TN* (true negatives), *FP* (false positives), and *FN* (false negatives) are fundamental metrics for evaluating model performance and form the basis for calculating other important indicators such as accuracy and recall.

- *TP* (true positives): the number of positive instances correctly predicted as positive.
- *TN* (true negatives): the number of negative instances correctly predicted as negative.
- *FP* (false positives): the number of negative instances incorrectly predicted as positive, also known as false alarms.
- *FN* (false negatives): the number of positive instances incorrectly predicted as negative, also known as missed detections.
- *Precision*: measures the overall proportion of correct predictions (both positive and negative) made by the model.
- *Recall*: measures the proportion of actual positives that are correctly identified by the model. While accuracy focuses on the overall correctness of the model's predictions, recall emphasizes the model's ability to capture positive instances.

For this study, there were three types of fruits in the experiment, so $k = 3$ in Formula (10). Within this context, mAP50 denotes the mAP (mean average precision) value of each category when the IoU (intersection over union) is set to 0.5, and mAP50-95 denotes the mAP on different IoU thresholds (ranging from 0.5 to 0.95). According to these indicators, we compared the PR (precision–recall) curves of the experimental data, as shown in Figure 9.

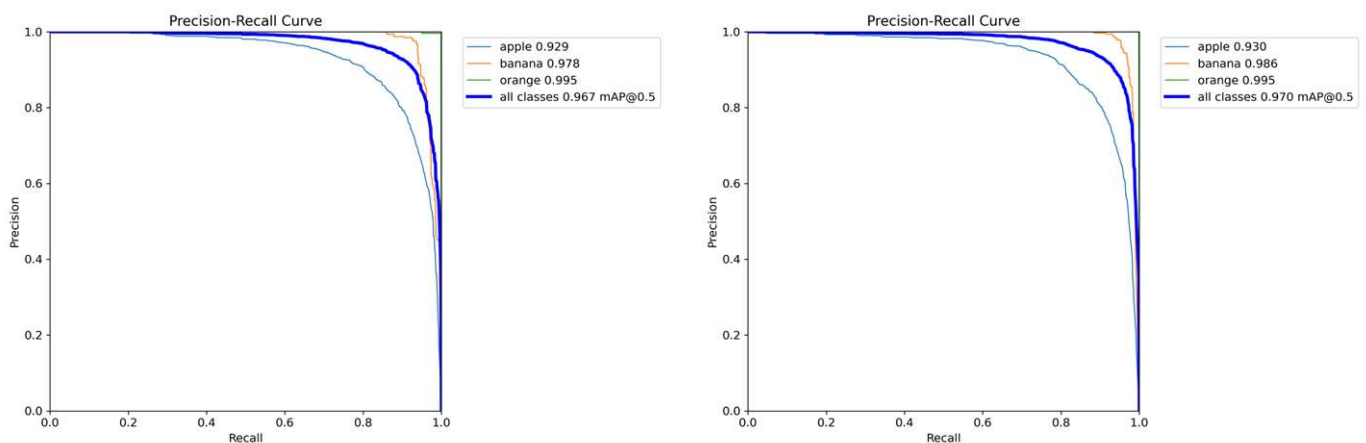


Figure 9. Comparison of P–R curves (left is the original YOLOv8 version, right is the improved version).

As can be seen, the improved model proposed in this study outperformed the original model. After training, TensorBoard was used to save the logs from both training and validation. The final performance metrics obtained by training with the original YOLOv8 are shown in Figure 10.

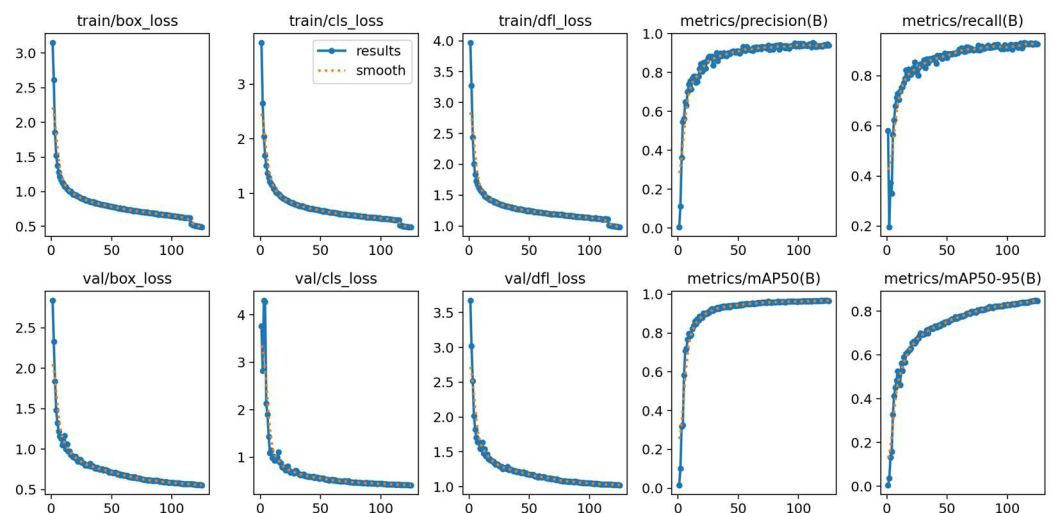


Figure 10. Performance metrics of the Original YOLOv8.

Following modifications like refining the loss function and integrating an attention mechanism, the mAP of the enhanced YOLOv8 neural network model had markedly increased relative to the original model. Specifically, the mAP50, which denotes the mAP value at an IoU threshold of 0.5, approached 0.970. The results of the training under the improved model are presented in Figure 11.

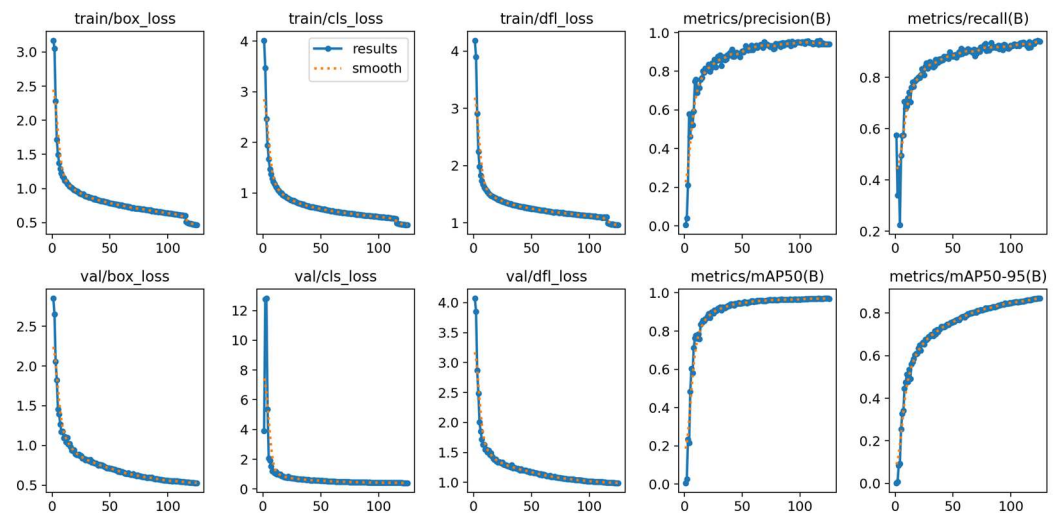


Figure 11. Performance graph of the improved YOLOv8.

Beyond the metrics traditionally utilized for assessing model performance, the training velocity of the models is also an essential factor to consider. Hence, subsequent to employing a pre-trained model and determining the optimal parameters across varying loss functions, this study incorporates FPS (frames per second) as a criterion to evaluate the computational speed of the model. FPS is calculated on a computer by employing the 'time()' function to capture the timestamps before and after processing a specified quantity of images, followed by dividing this duration by the total number of images processed. Figure 12 shows the recognition effect for fruits that move quickly. It can be seen that for fruits that moved quickly, even if there was blurring, the model was able to accurately recognize them. Thus, the model has good robustness.

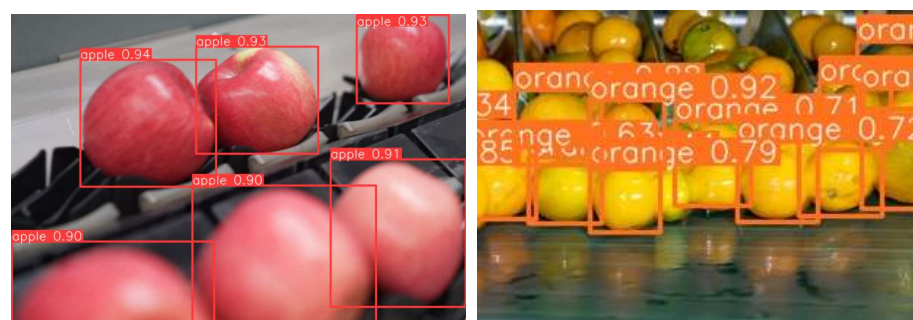


Figure 12. The effect of identifying fruits.

4.5. Comparative Experiment

To validate the effectiveness of the attention mechanism introduced in this paper, we conducted a series of comparative experiments. We compared several common attention mechanisms and explored the effects of the proposed improvements. We improved the network structure and optimized parameters of the YOLOv8n model, using the dataset from this study for training and validation. We compared the effectiveness of various improvement methods through experimental results. The main attention mechanisms compared included the SE (squeeze-and-excitation) attention mechanism, D-attention (vision

transformer with deformable attention), and LocalWindowAttention. Additionally, we included the YOLOv5 algorithm for comparison. The results of the comparative experiments are shown in Figure 13. The outcomes of these comparisons regarding the pertinent metrics are presented in Table 5.

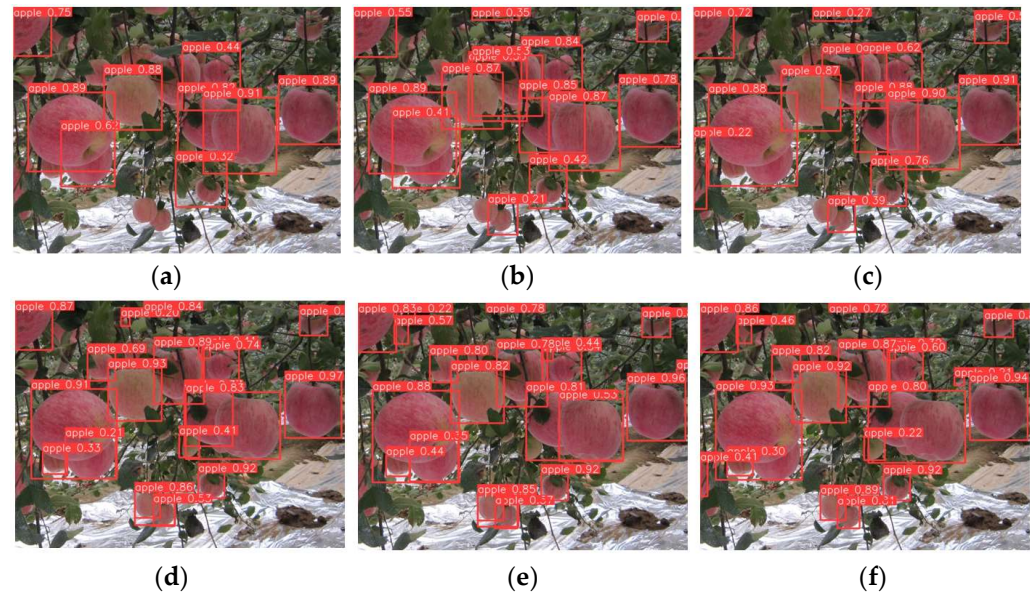


Figure 13. Detection results under small targets. (a) YOLOv5n; (b) YOLOv8n—LocalWindowAttention; (c) YOLOv8n—SEattention; (d) YOLOv8n—D-attention; (e) YOLOv8n; (f) Improved YOLOv8.

Table 5. Algorithm performance comparison.

Evaluating Indicator	Precision	Recall	mAP@0.5-0.95	FPS
YOLOv5n	0.88	0.88	0.82	180
YOLOv8n	0.91	0.90	0.84	202
YOLOv8n—SEattention	0.92	0.90	0.85	204
YOLOv8n—Dattention	0.90	0.91	0.84	205
YOLOv8n—LocalWindowAttention	0.87	0.89	0.82	198
Improved YOLOv8	0.93	0.92	0.86	212

As shown in Table 5, the precision–recall and mAP of the improved YOLOv8 were both higher than the other object detection methods, and its FPS value was also higher than the other three methods. Therefore, the improved YOLOv8 can quickly and accurately detect targets, and it meets industrial requirements. The results of object detection for the small target fruit dataset using the aforementioned four algorithms are shown in Figure 13.

Our improved algorithm’s detection results are shown in Figure 13f. It can be observed that the false negative rate of our improved attention mechanism was significantly lower than that of the YOLOv5n, YOLOv8n-LocalWindowAttention, and D-attention improved algorithms. Additionally, the detection accuracy was also superior to the algorithms based on SE-attention improvements, and the YOLOv8n algorithm itself.

At the same time, by comparing the impact of the commonly used loss functions IoU, DIoU (Distance-IoU), GIOU (Generalized-IoU), CIOU (Complete-IoU), and SIOU (SCYLLA-IoU) on model accuracy, we can see from the experimental results in Table 4 that the model using the SIOU loss function (YOLOv8n+SIOU) had higher accuracy. This indicates that the incorporation of this loss function effectively improved the mAP compared to the original model (YOLOv8n+CIoU). The results are shown in Table 6.

Table 6. Comparison of different loss functions.

Model	mAP@0.5/%	mAP@0.5-0.95/%	FPS
YOLOv8n+CIoU	96.7	84.6	202
YOLOv8n+DIoU	96.8	85.7	206
YOLOv8n+GIoU	96.5	84.2	208
YOLOv8n+SIOU	96.9	86.3	209

4.6. Ablation Experiment

To test the effectiveness of the improved model proposed in this study, we conducted comparative ablation experiments. These experiments were performed on the Fasternet module with the introduced attention mechanism (abbreviated as AFN), the small object detection layer (abbreviated as SODL), and the SIOU loss function. Where x represents not enabled, $\checkmark$ represents enabled. As can be seen from Table 7, by replacing the original c2f module with the improved attention module proposed in this study, adding the small object detection layer, and improving the loss function, the mAP @ 0.5 was effectively increased. Furthermore, by using the SIOU loss function, the limitations of CIoU in real small object detection scenarios were addressed, which had a certain enhancing effect on the detection performance of the model.

Table 7. Comparison results of different models for ablation experiments.

AFN	SODL	SIOU	AP50(%)	FPS
x	x	x	96.7	210
$\checkmark$	x	x	96.8	215
x	$\checkmark$	x	96.6	156
x	x	$\checkmark$	96.9	201
$\checkmark$	$\checkmark$	$\checkmark$	97.0	212

4.7. Heatmaps for Object Recognition

In the field of computer vision, heatmaps are a visualization tool used to display the intensity or importance of specific information, usually shown through changes in color. These maps typically use a color gradient from cold to warm colors, such as blue to red, where red generally represents higher values or larger focus points, and blue represents lower values or less focus.

In neural networks utilizing attention mechanisms, heatmaps aid in visualizing which regions of an image the model focuses on during processing. Heatmap generation also reveals the spatial support regions used for image classification decisions. This study utilized heatmaps generated by Grad-CAM (gradient-weighted class activation mapping) to identify which regions of an image contribute most significantly to classification results. These heatmaps highlight the regions activated in the network when performing specific tasks. One reason for choosing Grad-CAM is its capability to produce heatmap visualizations overlaid on the original image, enabling researchers and practitioners to intuitively interpret the model's behavior. This interpretability is crucial for understanding why the model makes certain predictions, particularly in critical domains like object detection. Moreover, compared to the original CAM, Grad-CAM is a gradient-based technique that does not require architectural changes or retraining of the model. It can be applied to any CNN-based architecture, making it versatile and applicable across various domains and tasks without significant overhead. A comparative analysis of the heatmaps generated using the original YOLOv8 model and the enhanced model is showcased in Figure 14.

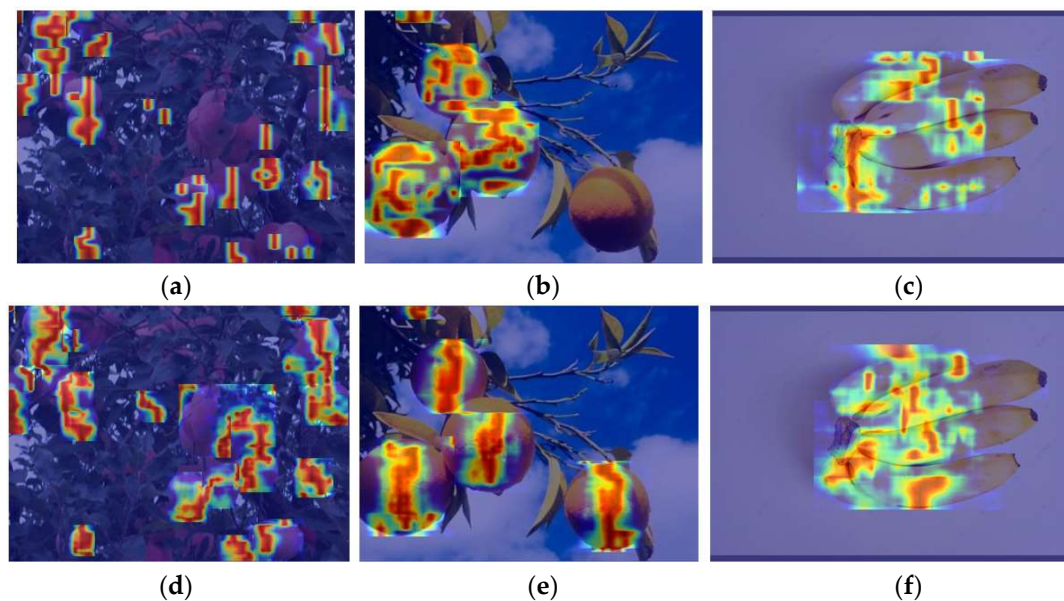


Figure 14. (a–c) represent the heatmaps of the original model identifying apples, oranges, and bananas, respectively. (d–f) represent the heatmap effects of our improved model identifying the same fruits.

4.8. Object Recognition Images

The prediction function of the enhanced YOLOv8 network model was utilized to evaluate the test set, with the results depicted in Figure 15. Within the previously described operational environment, the processing time for each image through the improved model was allocated as follows: 0.6 milliseconds for preprocessing, 95.6 milliseconds for inference, and 0.5 milliseconds for post-processing. Analyzing the displayed image alongside the prediction time reveals that the advanced YOLOv8 network is capable of identifying small target fruits both accurately and swiftly. As it has achieved good accuracy in multiple fruit category recognition tasks, the model has good generalization performance.

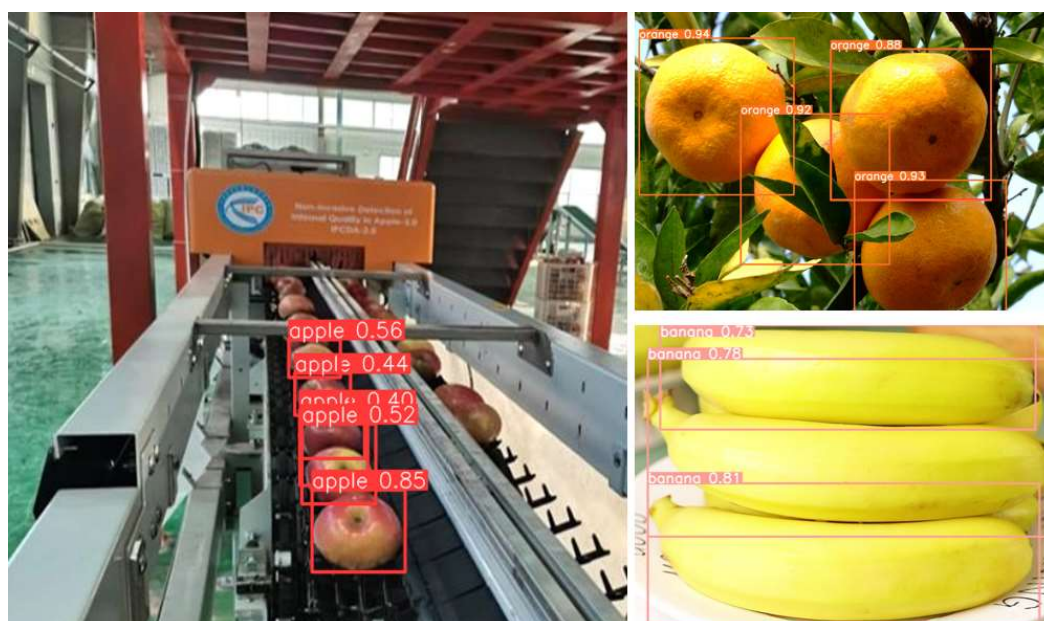


Figure 15. Recognition effect diagram of the improved YOLOv8 algorithm.

5. Conclusions

This paper improves the issues of low efficiency and high rate of missed detections and false pickups of a robotic arm when trying to identify and grasp small high-speed moving targets, based on the attention mechanism. The improved algorithm in this study not only enhances the accuracy and stability of tracking, but also enables the robotic arm to perform adaptive target tracking in complex and dynamic environments.

The method of machine vision is used in this study to allow a robotic arm to quickly grasp the target. The identification effect is significantly improved by modifying the YOLOv8 algorithm. The model's generalization ability is enhanced by means of image enhancement, while the introduction of the EMA attention mechanism greatly increases the model's success rate in recognizing moving small targets. In addition, local optima are avoided by improving the optimization of the loss function.

Finally, the experimental results on the small sample conveyor belt item dataset show that the improved YOLOv8 network proposed in this paper, based on the attention mechanism, can quickly and accurately identify targets. The statistical data of the experimental performance validate the effectiveness of the proposed method.

Author Contributions: Conceptualization, B.C. and A.J.; methodology, B.C.; software, Aishan jiang; validation, J.S., B.C. and A.J.; formal analysis, B.C.; investigation, A.J.; resources, J.S.; data curation, B.C.; writing—original draft preparation, A.J.; writing—review and editing, A.J.; visualization, A.J.; supervision, J.S.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Pagonis, K.; Zacharia, P.; Kantaros, A.; Ganetsos, T.; Brachos, K. Design, Fabrication and Simulation of a 5-Dof Robotic Arm Using Machine Vision. In Proceedings of the 2023 17th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania, 9–10 June 2023; IEEE: Oradea, Romania, 2023; pp. 1–4.
2. Jijesh, J.J.; Shankar, S.; Ranjitha; Revathi, D.C.; Shivaranjini, M.; Sirisha, R. Development of Machine Learning Based Fruit Detection and Grading System. In Proceedings of the 2020 International Conference on Recent Trends in Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 12–13 November 2020; IEEE: Bangalore, India, 2020; pp. 403–407.
3. Tan, H. Line Inspection Logistics Robot Delivery System Based on Machine Vision and Wireless Communication. In Proceedings of the 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Chongqing, China, 29–30 October 2020; IEEE: Chongqing, China, 2020; pp. 366–374.
4. Li, G.; Zhu, D. Research on Road Defect Detection Based on Improved YOLOv8. In Proceedings of the 2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 8–10 December 2023; IEEE: Chongqing, China, 2023; pp. 143–146.
5. Zhixin, L.; Yubo, H.; Tianding, Z.; Yueming, W.; Haoyuan, Y.; Wei, Z.; Yang, W. Discussion on the Application of Artificial Intelligence in Computer Network Technology. In Proceedings of the 2023 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS), Bristol, UK, 9–31 July 2023; IEEE: Bristol, UK, 2023; pp. 51–55.
6. Pedro, R.; Oliveira, A.L. Assessing the Impact of Attention and Self-Attention Mechanisms on the Classification of Skin Lesions. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–8.
7. Li, W.; Zhang, Z.; Li, C.; Zou, J. Small Target Detection Algorithm Based on Two-Stage Feature Extraction. In Proceedings of the 2023 6th International Conference on Software Engineering and Computer Science (CSECS), Chengdu, China, 22–24 December 2023; IEEE: Chengdu, China, 2023; pp. 1–5.
8. Singh, R.; Singh, D. Quality Inspection with the Support of Computer Vision Techniques. In Proceedings of the 2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC), Bengaluru, India, 18–19 November 2022; IEEE: Bengaluru, India, 2022; pp. 1584–1588.
9. Umanandhini, D.; Devi, M.S.; Beulah Jabaseeli, N.; Sridevi, S. Batch Normalization Based Convolutional Block YOLOv3 Real Time Object Detection of Moving Images with Backdrop Adjustment. In Proceedings of the 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, India, 17–19 August 2023; IEEE: Kochi, India, 2023; pp. 25–29.

10. Du, J.; Lu, H.; Zhang, L.; Hu, M.; Shen, X. Infrared Small Target Detection and Tracking Method Suitable for Different Scenes. In Proceedings of the 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 11–13 December 2020; IEEE: Chongqing, China, 2020; pp. 664–668.
11. Chen, X.; Guan, J.; Wang, Z.; Zhang, H.; Wang, G. Marine Targets Detection for Scanning Radar Images Based on Radar- YOLONet. In Proceedings of the 2021 CIE International Conference on Radar (Radar), Haikou, China, 5–19 December 2021; IEEE: Haikou, China, 2021; pp. 1256–1260.
12. Duth, S.; Vedavathi, S.; Roshan, S. Herbal Leaf Classification Using RCNN, Fast RCNN, Faster RCNN. In Proceedings of the 2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 18 August 2023; IEEE: Pune, India, 2023; pp. 1–8.
13. Wu, Z.; Yu, H.; Zhang, L.; Sui, Y. AMB: Automatically Matches Boxes Module for One-Stage Object Detection. In Proceedings of the 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Changchun, China, 11–13 August 2023; IEEE: Changchun, China, 2023; pp. 1516–1522.
14. Gai, R.; Li, M.; Chen, N. Cherry Detection Algorithm Based on Improved YOLOv5s Network. In Proceedings of the 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Haikou, China, 20–22 December 2021; IEEE: Haikou, China, 2021; pp. 2097–2103.
15. Pandey, S.; Chen, K.-F.; Dam, E.B. Comprehensive Multimodal Segmentation in Medical Imaging: Combining YOLOv8 with SAM and HQ-SAM Models. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2–6 October 2023; IEEE: Paris, France, 2023; pp. 2584–2590.
16. Gunawan, F.; Hwang, C.-L.; Cheng, Z.-E. ROI-YOLOv8-Based Far-Distance Face-Recognition. In Proceedings of the 2023 International Conference on Advanced Robotics and Intelligent Systems (ARIS), Taipei, Taiwan, 30 August–1 September 2023; IEEE: Taipei, Taiwan, 2023; pp. 1–6.
17. Samaniego, L.A.; Peruda, S.R.; Brucal, S.G.E.; Yong, E.D.; De Jesus, L.C.M. Image Processing Model for Classification of Stages of Freshness of Bangus Using YOLOv8 Algorithm. In Proceedings of the 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 10–13 October 2023; IEEE: Nara, Japan, 2023; pp. 401–403.
18. Shetty, A.D.; Ashwath, S. Animal Detection and Classification in Image & Video Frames Using YOLOv5 and YOLOv8. In Proceedings of the 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 22–24 November 2023; IEEE: Coimbatore, India, 2023; pp. 677–683.
19. Zhou, F.; Guo, D.; Wang, Y.; Zhao, C. Improved YOLOv8-Based Vehicle Detection Method for Road Monitoring and Surveillance. In Proceedings of the 2023 5th International Symposium on Robotics & Intelligent Manufacturing Technology (ISRIMT), Changzhou, China, 22–24 September 2023; IEEE: Changzhou, China, 2023; pp. 208–212.
20. Peri, S.D.B.; Palaniswamy, S. A Novel Approach To Detect and Track Small Animals Using YOLOv8 and DeepSORT. In Proceedings of the 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 6–8 October 2023; IEEE: Bangalore, India, 2023; pp. 1–6.
21. Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. MATNet: Motion-Attentive Transition Network for Zero-Shot Video Object Segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 8326–8338. [[CrossRef](#)]
22. Yang, H.; Lin, L.; Zhong, S.; Guo, F.; Cui, Z. Aero Engines Fault Diagnosis Method Based on Convolutional Neural Network Using Multiple Attention Mechanism. In Proceedings of the 2021 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), Weihai, China, 13–15 August 2021; pp. 13–18.
23. Luo, Z.; Li, J.; Zhu, Y. A Deep Feature Fusion Network Based on Multiple Attention Mechanisms for Joint Iris-Periocular Biometric Recognition. *IEEE Signal Process. Lett.* **2021**, *28*, 1060–1064. [[CrossRef](#)]
24. Shi, Y.; Hidaka, A. Attention-YOLOX: Improvement in On-Road Object Detection by Introducing Attention Mechanisms to YOLOX. In Proceedings of the 2022 International Symposium on Computing and Artificial Intelligence (ISCAI), Beijing, China, 16–18 December 2022; pp. 5–14.
25. Dong, Y. Research on Performance Improvement Method of Dynamic Object Detection Based on Spatio-Temporal Attention Mechanism. In Proceedings of the 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Kuala Lumpur, Malaysia, 8–11 October 2023; pp. 1558–1563.
26. Du, D.; Cai, H.; Chen, G.; Shi, H. Multi Branch Deepfake Detection Based on Double Attention Mechanism. In Proceedings of the 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), Changchun, China, 23–26 September 2021; pp. 746–749.
27. Liang, C.; Dong, J.; Li, J.; Meng, J.; Liu, Y.; Fang, T. Facial Expression Recognition Using LBP and CNN Networks Integrating Attention Mechanism. In Proceedings of the 2023 Asia Symposium on Image Processing (ASIP), Tianjin, China, 15–17 June 2023; pp. 1–6.
28. Wu, M.; Zhao, J. Siamese Network Object Tracking Algorithm Combined with Attention Mechanism. In Proceedings of the 2023 International Conference on Intelligent Media, Big Data and Knowledge Mining (IMBDKM), Changsha, China, 17–19 March 2023; pp. 20–24.
29. Yang, Y.; Sun, L.; Mao, X.; Dai, L.; Guo, S.; Liu, P. Using Generative Adversarial Networks Based on Dual Attention Mechanism to Generate Face Images. In Proceedings of the 2021 International Conference on Computer Technology and Media Convergence Design (CTMCD), Sanya, China, 23–25 April 2021; pp. 14–19.

30. Chen, C.; Wu, X.; Chen, A. A Semantic Segmentation Algorithm Based on Improved Attention Mechanism. In Proceedings of the 2020 International Symposium on Autonomous Systems (ISAS), Guangzhou, China, 6–8 December 2020; pp. 244–248.
31. Osama, M.; Kumar, R.; Shahid, M. Empowering Cardiologists with Deep Learning YOLOv8 Model for Accurate Coronary Artery Stenosis Detection in Angiography Images. In Proceedings of the 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 23–24 June 2023; pp. 1–6.
32. Wang, Z.; Luo, X.; Li, F.; Zhu, X. Lightweight Pig Face Detection Method Based on Improved YOLOv8. In Proceedings of the 2023 13th International Conference on Information Science and Technology (ICIST), Cairo, Egypt, 8–14 December 2023; pp. 259–266.
33. Gonthina, N.; Katkam, S.; Pola, R.A.; Pusuluri, R.T.; Prasad, L.V.N. Parking Slot Detection Using YOLOv8. In Proceedings of the 2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNBC), Tumkur, India, 8–10 December 2023; pp. 1–7.
34. Haimar, Z.; Mateur, K.; Farhan, Y.; Madi, A.A. Pothole Detection: A Performance Comparison Between YOLOv7 and YOLOv8. In Proceedings of the 2023 9th International Conference on Optimization and Applications (ICOA), Abu Dhabi, India, 5–6 October 2023; pp. 1–7.
35. Orchi, H.; Sadik, M.; Khaldoun, M.; Sabir, E. Real-Time Detection of Crop Leaf Diseases Using Enhanced YOLOv8 Algorithm. In Proceedings of the 2023 International Wireless Communications and Mobile Computing (IWCMC), Marrakesh, Morocco, 19–23 June 2023; pp. 1690–1696.
36. Tan, Y.K.; Chin, K.M.; Ting, T.S.H.; Goh, Y.H.; Chiew, T.H. Research on YOLOv8 Application in Bolt and Nut Detection for Robotic Arm Vision. In Proceedings of the 2024 16th International Conference on Knowledge and Smart Technology (KST), Krabi, Thailand, 28 February–2 March 2024; pp. 126–131.
37. Xie, S.; Chuah, J.H.; Chai, G.M.T. Revolutionizing Road Safety: YOLOv8-Powered Driver Fatigue Detection. In Proceedings of the 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Nadi, Fiji, 4–6 December 2023; pp. 1–6.
38. Afonso, M.H.F.; Teixeira, E.H.; Cruz, M.R.; Aquino, G.P.; Vilas Boas, E.C. Vehicle and Plate Detection for Intelligent Transport Systems: Performance Evaluation of Models YOLOv5 and YOLOv8. In Proceedings of the 2023 IEEE International Conference on Computing (ICOCO), Langkawi, Malaysia, 9–12 October 2023; pp. 328–333.
39. Afrin, Z.; Tabassum, F.; Kibria, H.B.; Imam, M.D.R.; Hasan, M.d.R. YOLOv8 Based Object Detection for Self-Driving Cars. In Proceedings of the 2023 26th International Conference on Computer and Information Technology (ICCIT), Toronto, ON, Canada, 13–15 December 2023; pp. 1–6.
40. Abyasa, J.; Kenardi, M.P.; Audrey, J.; Jovanka, J.J.; Justino, C.; Rahmania, R. YOLOv8 for Product Brand Recognition as Inter-Class Similarities. In Proceedings of the 2023 3rd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), Yogyakarta, Indonesia, 9–10 August 2023; pp. 514–519.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.