

Article

Public Security Patrol and Alert Recognition for Police Patrol Robots Based on Improved YOLOv8 Algorithm

Yuehan Shi, Xiaoming Zhang , Qilei Wang ^{*} and Xiaojun Liu

Faculty of Graduate Studies and Research, China People's Police University, Langfang 065000, China; 2023903033@ccpu.edu.cn (Y.S.); liuxiaojun@ccpu.edu.cn (X.L.)

^{*} Correspondence: zhangxiaoming@ccpu.edu.cn (X.Z.); wangqilei@ccpu.edu.cn (Q.W.); Tel.: +86-188-0019-9952 (X.Z.); +86-155-1264-6676 (Q.W.)

Abstract

Addressing the prevalent challenges of inadequate detection accuracy and sluggish detection speed encountered by police patrol robots during security patrols, we propose an innovative algorithm based on the YOLOv8 model. Our method consists of substituting the backbone network of YOLOv8 with FasterNet. As a result, the model's ability to identify accurately is enhanced, and its computational performance is improved. Additionally, the extraction of geographical data becomes more efficient. In addition, we introduce the BiFormer attention mechanism, incorporating dynamic sparse attention to significantly improve algorithm performance and computational efficiency. Furthermore, to bolster the regression performance of bounding boxes and enhance detection robustness, we utilize Wise-IoU as the loss function. Through experimentation across three perilous police scenarios—fighting, knife threats, and gun incidents—we demonstrate the efficacy of our proposed algorithm. The results indicate notable improvements over the original model, with enhancements of 2.42% and 5.83% in detection accuracy and speed for behavioral recognition of fighting, 2.87% and 4.67% for knife threat detection, and 3.01% and 4.91% for gun-related situation detection, respectively.

Keywords: YOLOv8; police patrol robot; security patrol; police recognition



Academic Editor: Leonardo Trujillo

Received: 18 July 2025

Revised: 6 September 2025

Accepted: 8 September 2025

Published: 10 September 2025

Citation: Shi, Y.; Zhang, X.; Wang, Q.; Liu, X. Public Security Patrol and Alert Recognition for Police Patrol Robots Based on Improved YOLOv8 Algorithm. *Math. Comput. Appl.* **2025**, *30*, 97. <https://doi.org/10.3390/mca30050097>

Copyright: © 2025 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

In the realm of law enforcement patrol tasks, the utilization of police robots presents a compelling alternative to human resources, offering all-weather functionality, efficiency, accuracy, and remote monitoring capabilities. This paper highlights the increasing significance of police robots in patrol work, contributing positively to community security and stability. Deployable in diverse settings such as streets, stations, airports, and large-scale events, police robots face a complex and dynamic urban environment characterized by numerous obstacles, including pedestrians, vehicles, and buildings, as well as fluctuating weather and lighting conditions. These challenges necessitate rapid data processing and precise target detection capabilities. To address these demands, efficient computer vision algorithms are employed to analyze real-time video data acquired during patrol operations. However, traditional You Only Look Once (YOLO) algorithms encounter limitations such as insufficient real-time processing, susceptibility to environmental factors, and suboptimal detection accuracy. This paper explores strategies for overcoming these challenges and enhancing the effectiveness of police patrol robots in ensuring public safety and security.

Security patrols often encounter challenging scenarios, such as violent altercations involving weapons, necessitating rapid and accurate detection methods. This paper addresses the complexities inherent in detecting such behaviors, including the diversity of actions, real-time decision-making requirements, occlusion-induced feature incompleteness, and increased misjudgments. To mitigate these challenges, Scientists have devised creative techniques.

Mostafa et al. [1] proposed a YOLO-based deep learning C-Mask model for real-time mask detection and recognition in public places through drone monitoring, greatly improving mask detection performance in terms of drone mobility and camera orientation adjustment. Zhou et al. [2] designed an efficient multitasking model that fully mines information, reduces redundant image processing calculations, and utilizes an efficient framework to solve problems such as pedestrian detection, tracking, and multi-attribute recognition, improving computational efficiency and decision-making accuracy. Wang et al. [3] applied object detection to the fields of security and counter-terrorism and proposed a Closed-Circuit Television (CCTV) autonomous weapon real-time detection method based on You Only Look Once version 4 (YOLOv4), which has greatly improved accuracy, real-time performance, and robustness. Alvaro et al. [4] proposed a dual recognition system based on license plate recognition and visual encoding in the field of recognition and detection. They evaluated the performance of both the public and proposed datasets using a multi-network architecture, making significant contributions to vehicle recognition. Azevedo et al. [5] combined the YOLOR-CSPNet (YOLOR-CSP) architecture with the DeepSORT tracker, which greatly improved inference efficiency. Pal et al. [6] proposed a new end-to-end vision-based detection, tracking, and classification architecture that enables robots to assist human fruit pickers in fruit picking operations. Ji et al. [7] proposed a small object detection algorithm called MCS-YOLOV4 based on YOLOv4 and the Corrected Intersection over Union (CIoU) loss function, which introduces an extended perception block domain and improves the attention mechanism and loss function, resulting in significant improvements in detecting small objects. Garcia-Cobo et al. [8] proposed a new deep learning architecture that uses Convolution Long Short-Term Memory (ConvLSTM), a convolutional alternative to standard Long Short-Term Memory (LSTM), to accurately and efficiently detect violent crimes in surveillance videos. In their research on violence detection, Rendón-Segador et al. [9] proposed the CrimeNet network, which combines a Vision Transformer (ViT) with a neural network trained with adversarial Neural Structured Learning (NSL) to improve detection accuracy.

The above mentioned algorithmic improvements have improved the target detection accuracy to different degrees; however, when detecting police situations such as fights, knives, guns, etc., there are still problems such as low real-time performance, difficulties in dealing with the situation of cover-up of characters, and frequent false detections, which lead to the accuracy and speed of police recognition not meeting the requirements. In order to improve the effect of the police patrol robot in the process of security patrol for the recognition of dangerous police situations, this paper improves the algorithm from the backbone network, the attention mechanism, and the loss function. Targeted optimization is carried out from the perspectives of real-time, accuracy and misjudgment difficulty, to improve its target detection performance and robustness in complex environments, and to provide a powerful and efficient solution for target detection of police patrol robots in security patrol scenarios.

2. Proposed Model

2.1. YOLOv8 Network

The backbone network in the design of YOLOv8 [10] plays a crucial role as the main component responsible for identifying features from the supplied image. Subsequently, the neck network undertakes the responsibility of integrating and refining the extracted features to facilitate classification and localization tasks. Ultimately, the head network is the crucial component of the target detection model, responsible for producing definitive detection outcomes. The network structure of YOLOv8 is schematically shown in Figure 1.

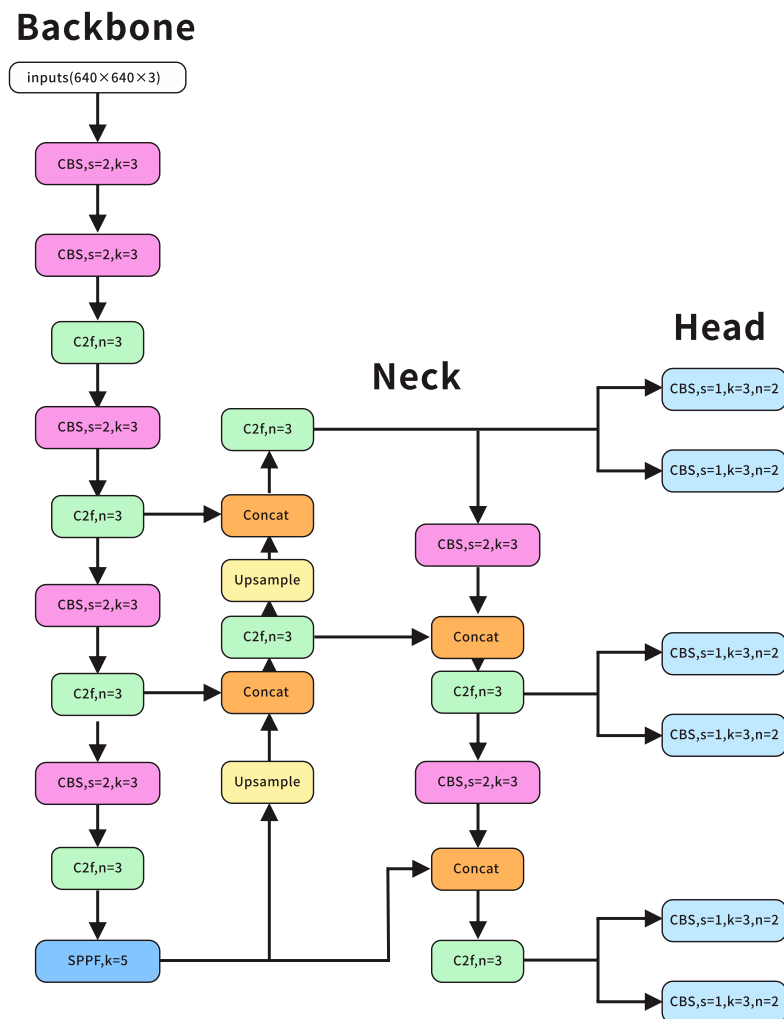


Figure 1. YOLOv8 network architecture.

YOLOv8 preserves the foundational network structure introduced in CSPDarknet within YOLOv5, while integrating the SPPF module from the YOLOv5 [11] architecture. Notably, the backbone network and neck segment adopt the C2f structure, inspired by the design philosophy of ELAN in YOLOv7 [12]. Leveraging the C2f structure enhances gradient flow through increased residual connections, thereby significantly expediting convergence. Furthermore, the detection head undergoes a transformation from a coupled head, as seen in YOLOv5, to a decoupled head configuration. This modification segregates the regression and prediction branches, resulting in expedited and enhanced convergence. Regarding the loss function, it encompasses two branches: classification and regression. The classification branch uses a simple binary cross-entropy (BCE) loss, whereas the regression branch uses a distribution focal loss. This helps the model converge faster by paying more attention to the distribution of nearby regions around the target position.

2.2. Improved YOLOv8 Network

In this study, our primary focus revolves around enhancing the YOLOv8 algorithm across three pivotal dimensions: the backbone network, attention mechanism, and loss function. To begin, we integrate FasterNet [13] in lieu of the conventional backbone network, a strategic maneuver aimed at augmenting computational efficiency while upholding superior accuracy, thereby realizing a lightweight design paradigm. Subsequently, BiFormer [14], an attention mechanism is used to enhance the accuracy of the model, thereby fortifying its resilience and detection precision. Finally, we employ the Wise Intersection Over Union (WIoU) [15] loss function, featuring a dynamic non-monotonic focusing mechanism, to optimize the model's performance concerning loss function dynamics.

2.2.1. Backbone Network Improvements

MobileNet [16], ShuffleNet [17], and GhostNet [18] leverage deep convolution (DW-Conv) or group convolution (GConv) methodologies to extract spatial features, thereby diminishing latency through a reduction in floating-point operations while inevitably increasing memory accesses. Conversely, MicroNet endeavors to decompose the network, thereby reducing FLOPs; however, its computational fragmentation proves inefficient, resulting in frequent memory accesses, particularly evident in deep separable convolutions, leading to decreased FLOPs. In response, the concept of partial convolution (PConv) emerged, aiming to concurrently mitigate computation, redundancy, and memory access. In 2023, Chen et al. [19] introduced FasterNet, building upon the principles of PConv with a focus on lightweight design, efficient feature extraction, and an optimized anchor mechanism. The working schematic is shown in Figure 2.

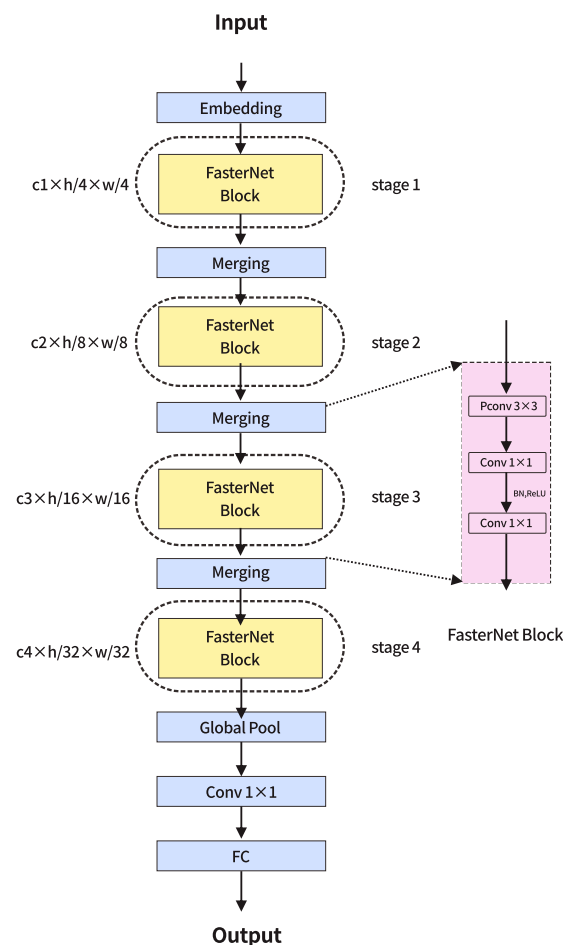


Figure 2. FasterNet architecture.

The PConv approach exploits redundancy in the feature map by selectively applying normal convolution (Conv) to a portion of the input channels, while leaving the rest unchanged. This approach ensures that only specific channels undergo convolution, preserving the integrity and structure of the overall feature map. The model delay is computed as follows.

$$\text{Latency} = \frac{FLOPs}{FLOPS'} \quad (1)$$

FLOPs represent the aggregate count of floating-point operations (FPOs) undertaken by a computational model, whereas FLOPS denotes the rate of floating-point operations executed per second, serving as a metric to gauge computational efficiency. In contrast to conventional convolution, PConv exhibits diminished FLOPs yet augmented FLOPS. Given that heightened FLOPs correspond to reduced latency within the model, the significance of FLOPS amplifies accordingly. The formula for calculating FLOPs for PConv is:

$$FLOPs = h \times w \times k^2 \times c_p^2. \quad (2)$$

The memory access for PConv can be described by the following equation, where h and w represent the width and height of the feature map, k denotes the size of the convolution kernel, and c_p is the number of channels on which the conventional convolution operates. In practical implementations, there are typically M input channels and N output channels, resulting in a reduction in FLOPs to just 1/16 of that achieved by conventional convolution. The memory access of PConv is:

$$\text{Memory Access} = h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p. \quad (3)$$

The utilization of PConv results in a substantial reduction in memory accesses, amounting to only 25% of those necessitated by regular convolution. Notably, the unutilized channels, represented by (c_p) , remain untouched during computation, thus obviating the need for memory accesses.

Accelerated training velocity not only conserves time and resources but also expedites the iterative and optimization phases of model development. PConv enhances network efficiency by curtailing computational load and memory access through selective processing of input channels. This approach, markedly quicker and more precise compared to conventional convolutional techniques, facilitates FasterNet in achieving heightened computational velocity while upholding model accuracy.

2.2.2. Attention Mechanisms

Attention, a pivotal component within the Vision Transformer (VT) architecture, plays a crucial role in capturing extensive dependencies across input sequences through the self-attention mechanism. This functionality, while bringing some degree of performance improvement, increases the computational burden and takes up a large amount of memory. The original implementation of attention, which is directly global in scope, leads to high computation and large memory footprint. This problem was later addressed by introducing content-independent sparsity to the attention mechanism.

The BiFormer model utilizes a bi-level routing mechanism (Bi-level routing) and Dynamic sparse attention to optimize the allocation of computational resources, enabling the capture of global and local features while maintaining flexibility. When processing images, key-value pairs that are irrelevant to the task at hand can be effectively filtered out, focusing only on those regions that are critical for decision-making. The model applies fine-grained token-to-token attention within these selected regions. To achieve this, BiFormer constructs a directed graph using an adjacency matrix that reveals the

complex relationships between different key-value pairs. In this way, the model is able to identify the necessary regions for each particular region, thus ensuring efficient utilization of computational resources. Finally, this fine-grained attention mechanism is implemented through a region-to-region routing index matrix.

In the architecture of BiFormer, an innovative approach utilizing overlapping block embedding is employed to efficiently process the input image. Subsequently, from the second to the fourth stage, block merging modules will be employed to efficiently reduce the spatial resolution while expanding the channel depth, thus enriching the feature representation. In these phases, successive BiFormer blocks are transformed with features. At the beginning of each BiFormer block, deep convolution is employed to imply relative position information in the image. The model seamlessly integrates the Bidirectional Relational Attention (BRA) module and the Multi-Layer Perception (MLP) module. The BRA module coordinates the cross-positional relational modeling, while the MLP module performs the detailed processing of the per-position embedding information, which enhances the model's acuity in discriminating local features. With these constituent elements and processing, BiFormer achieves superior performance in a wide range of vision tasks, providing a powerful and novel tool for research and practical applications in the field of computer vision. The architecture diagram of BiFormer is shown in Figure 3.

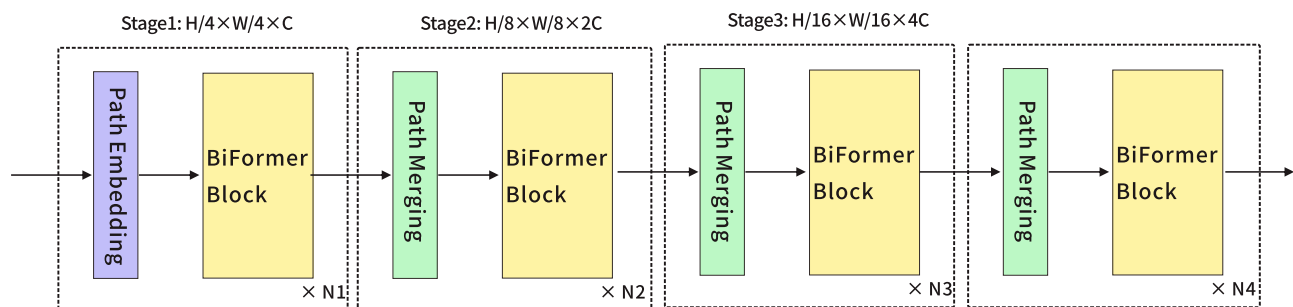


Figure 3. BiFormer Architecture.

2.2.3. Optimizing the Loss Function

Target detection is a crucial task in the field of computer vision, heavily reliant on the intricacies of the loss function for both detection accuracy and efficiency. Among its constituents, the bounding box loss function emerges as a linchpin in enhancing model performance. Current research efforts have mostly focused on improving the accuracy of this bounding box loss function. Nevertheless, in practical scenarios, training datasets often encompass instances of low-quality examples, which can detrimentally impact model training. Excessively prioritizing bounding box regression on such low-quality instances may impede overall model enhancement. Consequently, it becomes imperative to contemplate strategies for augmenting its efficacy in accommodating high-quality examples while mitigating the adverse effects of low-quality instances on model training. Introducing novel mechanisms or strategies to strike a finer balance among the disparate qualities of examples during the training regimen holds promise for fostering more steadfast and efficient target detection outcomes.

To tackle the issues outlined earlier, Focal-EIoUv1 [20] was initially suggested. However, its fixed focusing mechanism did not fully exploit the capabilities of the non-monotonic focusing mechanism. In light of this limitation, a dynamic non-monotonic focusing mechanism is advanced herein, accompanied by the design of WIoU. This novel approach enhances target detection accuracy by integrating a weighting factor and considering the shape characteristics of overlapping regions within the loss function. By accentuating the degree of overlap and shape resemblance of target-bounding boxes, this

strategy facilitates improved model learning and adaptation to diverse shapes and sizes of target boxes. Furthermore, a distance attention (DA) mechanism, grounded in a distance metric, is introduced to augment the model's focus not only on target box position and size but also on the relative distance relationships between boxes during training. Such a meticulously crafted approach contributes to heightened detection performance and generalization capabilities. Therefore, this article introduces a WIoU loss function with a dynamic non monotonic focusing mechanism to optimize the model. By replacing Intersection over Union (*IoU*) with outlier degree through the dynamic non monotonic focusing mechanism to evaluate the quality of anchor boxes, the excessive punishment of geometric factors on the model is weakened, and the intervention in model training is reduced to increase the generalization ability of the model. That is:

$$IoU = \frac{W_i H_i}{S_u}, \quad (4)$$

$$\mathcal{L}_{IoU} = 1 - IoU, \quad (5)$$

$$\mathcal{R}_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)}\right), \quad (6)$$

$$\mathcal{L}_{WIoUv1} = \mathcal{R}_{WIoU} \mathcal{L}_{IoU}. \quad (7)$$

x and y are the center point coordinates of the prediction box; x_{gt} and y_{gt} are the center point coordinates of the real box. W_g and H_g are the width and height of the minimum bounding box between the predicted box and the real box, W_i and H_i are the width and height of the intersection area between the predicted box and the real box, and S_u is the joint area between the predicted box and the real box. $\mathcal{L}_{IoU}$ represents *IoU* loss. $\mathcal{R}_{WIoU}$ is a distance metric function used to maintain good generalization ability of the model. $\mathcal{L}_{WIoUv1}$ is the loss function of *WIoUv1* (Wise Intersection Over Union Version 1). The schematic diagram of the predicted box and the real box is shown in Figure 4. To prevent $\mathcal{R}_{WIoU}$ from generating gradients that hinder convergence, W_g and H_g are separated from the computational graph. Because it effectively eliminates factors that hinder convergence, no new metrics such as aspect ratio are introduced. Thus, it reduces the impact of aspect ratio on the convergence of the original loss function. The red connecting line represents the distance between the predicted box center point and the real box center point. The light green area is the overlapping area between the predicted box and the real box in the graph.

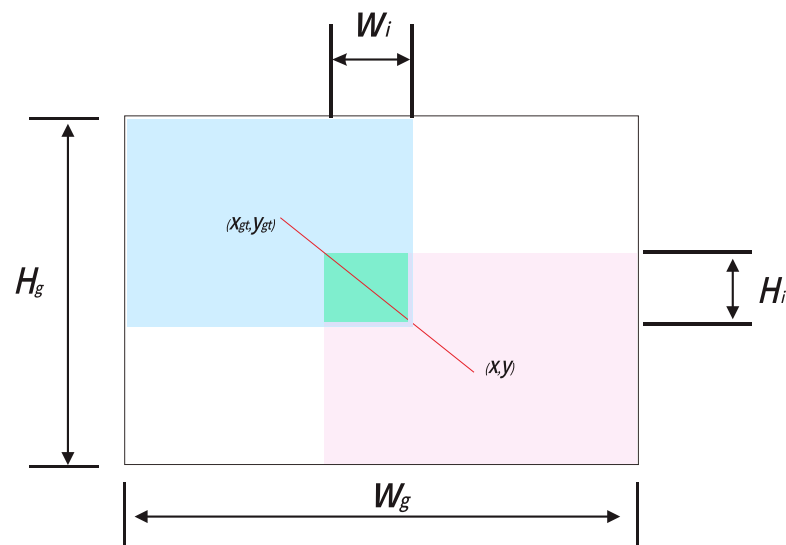


Figure 4. Schematic diagram of the prediction box and real box.

3. Experiments

3.1. Experimental Environment and Dataset

The experimental environment configuration is shown in Table 1.

Table 1. Configuration of experimental environment.

Form	Version Number
Pytorch	Torch-1.13.1
CPU	Intel Core i5-10400F @ 2.9 GHz
GPUs	RTX3080 (24G)
Operating system	Windows 10
Programming language	Python 3.8

In terms of hardware, the system is installed on the police patrol robot platform, which is developed and modified using the RCBOT-PR-01 robot developed by Suzhou Rongcui Special Robot Co., Ltd. in Suzhou, China. The platform is equipped with a binocular gimbal camera, laser radar, ultrasonic wave sensors, etc. The upper computer of the platform is a laptop computer.

Public datasets such as Hockey Fight Detection Dataset [21], ImageNet [22], Open Images Dataset [23], SVHN [24] are used to collect images and videos related to dangerous behaviors such as fights, knives and guns. After obtaining the public datasets, the video data is processed by frame splitting to eliminate redundant information, reduce duplicate images, and avoid the overfitting phenomenon caused by too many duplicate images. Adopting the method of extracting 1 frame for every 10 frames, a total of 5987 video images of fighting behavior, 6878 video images of gun-wielding behavior, and 6028 video images of knife-wielding behavior are acquired in this process.

In addition, 498 pictures of real-life fighting and assault behaviors, 526 pictures of knife-wielding behaviors, and 325 pictures of gun-wielding behaviors were captured by the camera. Finally, after screening, 5002 pictures of fighting and assault behavior, 5257 pictures of knife-wielding behavior, and 5029 pictures of gun-wielding behavior were selected for the dataset, and all pictures were renamed in .jpg format. Subsequently, the labelling (version 1.3.3) was used to manually label the violent behavior pictures, which provided diversified data for the subsequent detection and identification of the behaviors. Finally, the dataset was divided into training set, validation set and test set in the ratio of 8:1:1. There are 3979 images in the training set, 519 images in the validation set and 504 images in the test set. The number of samples of the divided images is shown in Table 2.

Table 2. Sample size.

Behavioral Segmentation	Training Set (Train)	Validation Set (val)	Test Set (Test)
Fighting dataset	3979	519	504
Knife dataset	4234	500	523
Gun dataset	4062	478	489

3.2. Evaluation Indicators

In this paper, Accuracy (Precision), Recall, Mean Accuracy (mAP), Mean Average Precision at $IoU = 0.5$ (mAP@0.50), Model Parameters, Model Complexity (GFLOPs) and Frames per Second (FPS) in the YOLOv8 algorithm are selected as the performance evaluation metrics to evaluate the performance of the target detection algorithm. The calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP}, \quad (8)$$

$$Recall = \frac{TP}{TP + FN}, \quad (9)$$

where TP denotes a positive sample that is predicted by the model to be a positive class, FP denotes a negative sample that is predicted by the model to be a positive class, and FN denotes a positive sample that is predicted by the model to be a negative class.

$$AP = \int_0^1 P(R) dR, \quad (10)$$

$$mAP = \frac{\sum_{m=1}^M AP_m}{M}, \quad (11)$$

AP is the average of Precision (P) at different Recall rates (Recall). After that the AP of all categories are averaged to obtain mAP .

$$FPS = \frac{N(p)}{T(p)} \quad (12)$$

$N(p)$ denotes the total number of images processed and $T(p)$ denotes the time taken to process (p) the images.

The number of model parameters (Params) is the sum of all parameters in the model and is related to the generalization ability of the model. Model complexity (GFLOPs) is used to measure the computational speed and capability of a neural network model, with higher values of GFLOPs usually indicating that the model is able to process data faster. $mAP@0.50$ is a key evaluation indicator in the field of object detection, representing the average accuracy of multiple categories calculated at an IoU threshold of 0.5.

3.3. Analysis of Experimental Results

3.3.1. Comparison of Backbone Network Improvements

ShuffleNetV2 [25] was proposed by the team of Kuangwei Technology in 2018, using direct metrics to replace indirect metrics and evaluated on the target platform, as can be seen from the table, it carries out a small enhancement in detection speed with the help of a lightweight network structure compared to the CSPDarknet structure, but there is a certain degree of reduction in recall and accuracy. GhostNet employs fewer parameters to generate more feature maps with reduced computational complexity, and its higher value of GFLOPs reflects the ability to process data at high speed, but $mAP@0.50$ Decreased by 0.11 percentage points. VanillaNet [26] removes the residuals and the attention module and employs the simplest convolutional computation, which provides an improvement of 1.4% in terms of accuracy, $mAP@0.50$ Indicator increased by 2.51%. EfficientNetV2 [27] further improves the detection speed compared to the CSPDarknet structure by using a combination of training perceptual neural structure search and scaling, and it has slightly increased $mAP@0.50$ and in terms of model complexity.

FasterNet improves the detection speed to a greater extent by utilizing its partial convolution compared to the original structure CSPDarknet, and the $mAP@0.50$ improved indicators by 3.05%, and the results are shown in Table 3.

Table 3. Comparative experimental results of backbone network. The bold numbers denote the best results.

Backbone Network	Precision/%	Recall/%	mAP@0.50/%	Params/ 10^6	GFLOPs	FPS
CSPDarknet	86.13	86.37	86.12	31.57	8.9	175.33
ShuffleNetV2	85.58	86.11	87.42	63.88	16.5	180.40
GhostNet	87.08	87.05	86.01	28.21	7.8	175.02
VanillaNet	87.53	86.06	88.61	17.34	8.1	188.92
EfficientNetV2	87.64	86.05	86.80	21.26	9.0	177.32
FasterNet	88.13	86.29	89.17	30.56	9.5	198.28

3.3.2. Comparison of Attention Mechanisms

In order to improve the detection accuracy, attention mechanisms are added to the original YOLOv8 for comparison experiments. Attention mechanisms such as ECA [28], CBAM [29], CA [30] and BiFormer attention mechanism selected in this paper are chosen for comparison experiments. The Efficient Channel Attention (ECA) module, based on deep CNNs, is a method that effectively captures cross-channel interactions without reducing dimensionality. Experimental validation on the dataset shows that using the ECA module improves the mAP@0.50 metric by 0.87%. CBAM incorporates channel attention and spatial attention mechanisms to improve the capacity of various channel characteristics to capture and extract important spatial information at different positions in the space. This results in a 0.27% enhancement in its mAP@0.50 performance. The flexible and lightweight CA attention mechanism improves on the channel information and orientation-dependent location information, it is mAP@0.50 improved by 0.38%. The experimental results are shown in Table 4.

Table 4. Results of comparative experiments on attention mechanism. The bold numbers denote the best results.

Attention Mechanism	Precision/%	Recall/%	mAP@0.50/%	Params/ 10^6	GFLOPs	FPS
-	86.13	86.37	86.12	31.57	8.9	175.33
ECA	86.99	87.33	86.99	31.31	5.9	176.38
CBAM	86.39	88.10	86.39	30.29	6.0	176.31
CA	86.10	88.25	86.50	30.25	6.1	177.03
BiFormer	86.32	88.03	87.32	31.39	5.7	177.02

The addition of the BiFormer attention mechanism resulted in enhanced identification accuracy for all three datasets. The mAP@0.50 metric showed an improvement of 1.2%. Additionally, precision and recall increased by 0.21 and 1.66 percentage points, respectively.

3.3.3. Comparison of Loss Functions

This paper analyzes and compares experimental data of the original YOLOv8 algorithm and the improved YOLOv8 algorithm that incorporates the Wise-IoU loss function. The analysis is based on Figure 5, which displays the loss function curves. The blue solid line represents the loss function curve of the original YOLOv8 algorithm, while the red solid line represents the curve of the improved loss function. Compared to the other three algorithms, precision has been improved, Params is also smaller, model complexity is lower, and performance needs to be improved to some extent. Table 5 shows the comparative data of various loss functions. We can see that the combination of YOLOv8n and WIoU loss functions shows better performance in Precision, Recall, and Map metrics compared to Generalized-IoU [31] and Distance-IoU [32].

Table 5. Results of comparative experiments on loss function. The bold numbers denote the best results.

Backbone Network	Precision/%	Recall/%	mAP@0.50/%	Params/ 10^6	GFLOPs	FPS
YOLOv8n	86.13	86.37	86.12	31.57	8.9	175.33
YOLOv8n+GIoU	87.35	85.37	82.16	30.78	8.9	178.22
YOLOv8n+DIoU	87.29	86.05	89.13	32.10	9.0	173.32
YOLOv8n+WIoU	88.13	86.29	89.17	30.56	9.5	175.28

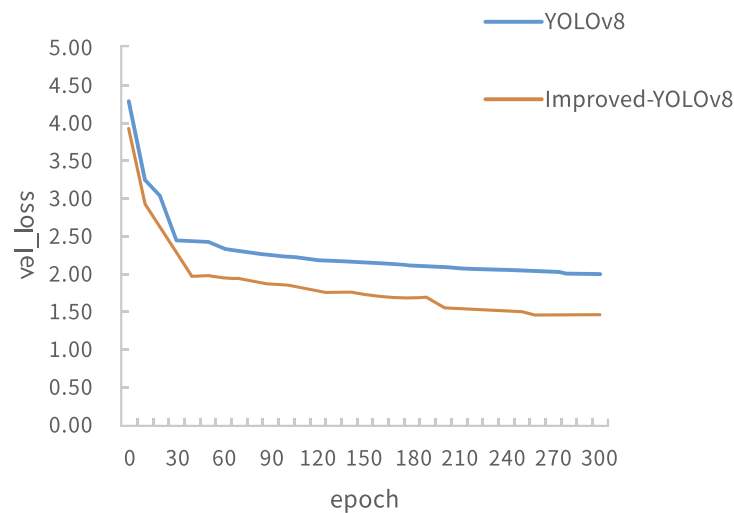


Figure 5. Loss curve graph before and after improvement.

3.3.4. Ablation Experiments

By replacing the backbone network of the original YOLOv8 algorithm with FasterNet, mAP@0.50, precision has been improved by 3.05 and 2 percentage points, respectively, and the FPS metrics have been improved to a greater extent. By adding BiFormer attention mechanism, it is mAP@0.50 improved by 1.2%. By the replacement of Wise-IoU loss function, it is mAP@0.50 improved by 0.4%

The experimental results presented in Table 6 demonstrate that simultaneous implementation of the three proposed enhancements in this study significantly enhances the performance of the algorithm in comparison to the original YOLOv8 method. The mAP@0.50 metric has been enhanced by 2.42%. Additionally, the GFLOPs and FPS have improved by 0.5 and 10.23, respectively. The accuracy and recall have also seen improvements of 1.41% and 2.61%, enabling a better level of precision to be achieved. Minimizing computational requirements and reducing data processing time significantly enhances algorithm performance, achieving lightweighting objectives without compromising computational accuracy.

Table 6. Results of ablation experiments. The bold numbers denote the best results.

FasterNet	BiFormer	Wise-IoU	Precision/%	Recall/%	mAP@0.50/%	Params/ 10^6	GFLOPs	FPS
			86.13	86.37	86.12	31.57	8.9	175.33
✓			88.13	86.29	89.17	30.56	9.5	198.28
	✓		86.32	88.03	87.32	31.39	5.7	177.02
		✓	85.60	86.25	86.52	30.25	9.1	178.03
✓	✓		86.98	89.10	87.28	31.29	9.3	192.08
	✓	✓	87.32	88.03	87.72	31.39	8.6	177.02
✓		✓	86.99	88.12	87.88	30.12	9.2	189.43
✓	✓	✓	87.54	88.98	88.54	31.18	9.4	185.56

3.3.5. Comparative Experiments with Different Models

The algorithm has been enhanced based on the original YOLOv8 model. This work also validates the method and its modules through comparative tests, while ensuring that each trial setting remains unaltered. The experimental findings may be seen in Table 7. When comparing the new algorithm to other YOLO models like YOLOv4, YOLOv5, and YOLOv8, the improved algorithm not only surpasses them in accuracy, but also effectively manages the number of parameters and computation required. It is well-suited for accurately detecting police situations. Figure 6 illustrates the comparison between the detection

performance of the enhanced YOLOv8 algorithm and the original YOLOv8 method on the dataset containing instances of fighting and assault.

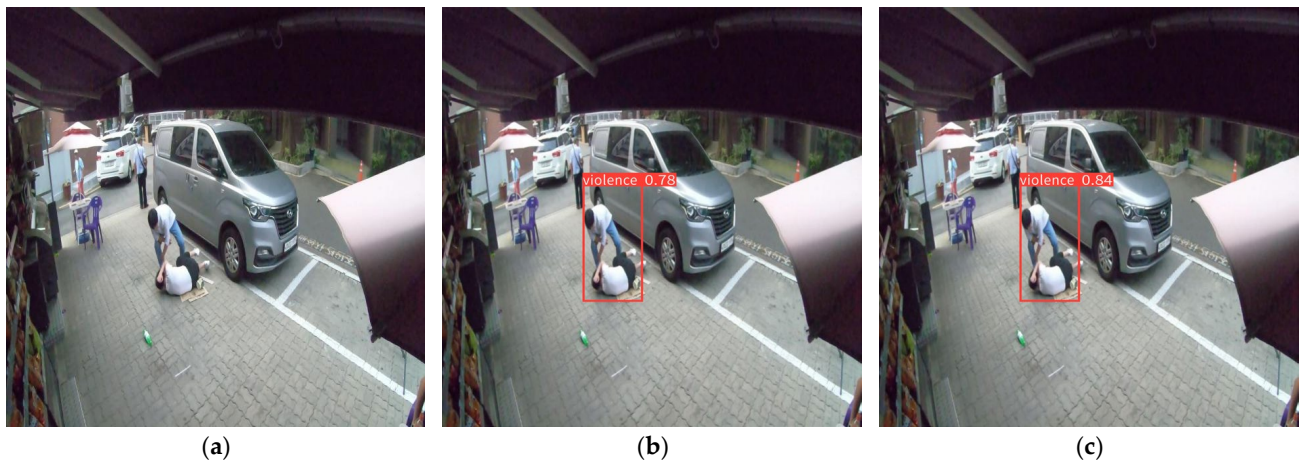


Figure 6. Comparison of detection effect (fight datasets). (a) Original (b) YOLOv8 (c) Improved YOLOv8.

Figures 6–8 illustrates the comparison between the detection performance of the enhanced YOLOv8 algorithm and the original YOLOv8 method on three publicly available datasets.

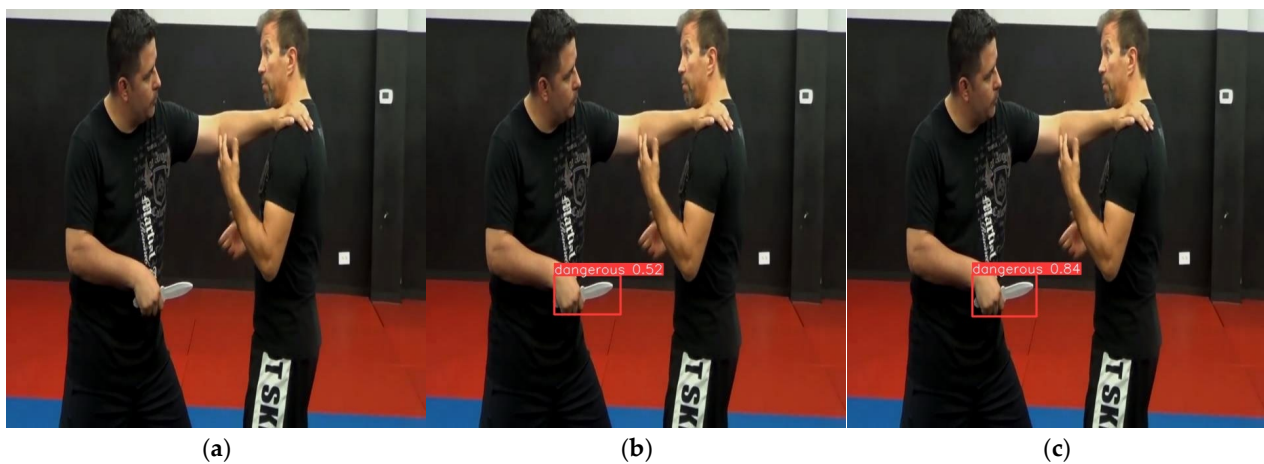


Figure 7. Comparison of detection effects (knife datasets). (a) Original (b) YOLOv8 (c) Improved YOLOv8.



Figure 8. Comparison of detection effects (gun datasets). (a) Original (b) YOLOv8 (c) Improved YOLOv8.

Table 7. Results of comparative experiments. The bold numbers denote the best results.

Model	Precision/%	Recall/%	mAP@0.50/%	Params/ 10^6	GFLOPs	FPS
YOLOv8n	86.13	86.37	86.12	31.57	8.9	175.33
YOLOv4	85.34	86.02	83.13	30.32	5.3	149.15
YOLOv5s	85.99	87.33	84.19	17.31	6.9	158.38
YOLOv5n	85.81	88.10	84.23	17.29	6.0	164.31
YOLOv8s	86.60	88.25	86.10	18.25	7.9	179.03
YOLOv8m	86.38	89.10	86.28	17.29	7.3	180.08
Improved-YOLOv8	87.54	88.98	88.54	31.18	9.4	185.56

4. Conclusions

In large security events, airport lobbies, station waiting rooms, and other densely populated areas with high crowd mobility, the target person may be partially or completely obscured, and dangerous behavior may occur in an instant. Robots are required to have ultra-high frame rate shooting and real-time processing capabilities. In addition, criminals may intentionally hide weapons or use crowd cover to make it difficult for robots to capture key frames. The improved model achieves high efficiency and accuracy in identifying dangerous behaviors such as holding knives and guns during the movement of police patrol robots in densely populated areas.

The enhanced algorithm suggested in this research employs the lightweight FasterNet network to substitute the initial backbone network. This substitution enhances the detection speed of the model while ensuring the accuracy of detection. Additionally, the inclusion of the attention mechanism BiFoemer further enhances the model's accuracy. Moreover, the utilization of the WIoU as a loss function results in the enhancement of the bounding box regression performance of the detection model. The dataset for this work was created by combining the publicly accessible dataset with the pertinent image videos captured during real-life police incidents.

The experimental results of the improved model on the dataset show that in the fighting and assault dataset, compared with the original YOLOv8, in the two evaluation indexes of average detection accuracy and detection speed, the detection of fighting and assault behavior recognition has been improved by 2.42% and 5.83%, respectively, the detection of knife-wielding behavior recognition has been improved by 2.87% and 4.67%, respectively, and the detection of gun-wielding behavior recognition has been improved by 3.01% and 4.91%, respectively. It is capable of performing the target detection task of police situation recognition in the process of security patrol. In the future, we will continue to optimize the model on the basis of the model, so that the model has higher detection ability and faster detection speed when facing more complex security patrol scenes and police situations.

At present, police patrol robots have made significant improvements in terms of functionality, but they still need to undergo practical testing when dealing with complex scenarios. In addition, in terms of laws and regulations, it remains to be further clarified whether police patrol robots can serve as independent law enforcement entities and whether law enforcement procedures comply with current legal provisions.

Author Contributions: Conceptualization, Q.W.; methodology, Y.S.; software, X.Z.; validation, X.Z.; formal analysis, Y.S.; investigation, X.L.; resources, Y.S., X.Z., Q.W. and X.L.; data curation, Q.W.; writing—original draft preparation, Y.S.; writing—review and editing, Y.S.; visualization, Q.W.; supervision, X.L.; project administration, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Equipment Research and Development Plan—Research on Key Technologies and Equipment of High Mobility Intelligent Unmanned Aerial Vehicles in the Perspective of Water Area ‘Double Anti’ Mission (2023ZB03).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

- Mostafa, S.A.; Ravi, S.; Zebari, D.A.; Zebari, N.A.; Mohammed, M.N.; Nedoma, J.; Martinek, R.; Deveci, M.; Ding, W. A YOLO-based deep learning model for Real-Time face mask detection via drone surveillance in public spaces. *Inf. Sci.* **2024**, *676*, 120865. [\[CrossRef\]](#)
- Zhou, Y.; Zeng, X.J.R. Towards comprehensive understanding of pedestrians for autonomous driving: Efficient multi-task-learning-based pedestrian detection, tracking and attribute recognition. *Robot. Auton. Syst.* **2024**, *171*, 104580. [\[CrossRef\]](#)
- Wang, G.; Ding, H.; Duan, M.; Pu, Y.; Yang, Z.; Li, H. Fighting against terrorism: A real-time CCTV autonomous weapons detection based on improved YOLO v4. *Digit. Signal Process.* **2023**, *132*, 103790. [\[CrossRef\]](#)
- Li, J.; Li, R.; Li, J.; Wang, J.; Wu, Q.; Liu, X. Dual-view 3d object recognition and detection via lidar point cloud and camera image. *Robot. Auton. Syst.* **2022**, *150*, 103999. [\[CrossRef\]](#)
- Azevedo, P.; Santos, V. Comparative analysis of multiple YOLO-based target detectors and trackers for ADAS in edge devices. *Robot. Auton. Syst.* **2023**, *171*, 104558. [\[CrossRef\]](#)
- Pal, A.; Leite, A.C.; From, P.J. A novel end-to-end vision-based architecture for agricultural human–robot collaboration in fruit picking operations. *Robot. Auton. Syst.* **2023**, *172*, 104567. [\[CrossRef\]](#)
- Ji, S.-J.; Ling, Q.-H.; Han, F. An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information. *Comput. Electr. Eng.* **2022**, *105*, 108490. [\[CrossRef\]](#)
- Garcia-Cobo, G.; SanMiguel, J.C. Human skeletons and change detection for efficient violence detection in surveillance videos. *Comput. Vis. Image Underst.* **2023**, *233*, 103739. [\[CrossRef\]](#)
- Rendón-Segador, F.J.; Álvarez-García, J.A.; Salazar-González, J.L.; Tommasi, T. CrimeNet: Neural Structured Learning using Vision Transformer for violence detection. *Neural Netw.* **2023**, *161*, 318–329. [\[CrossRef\]](#)
- Vazquez, F.; Nuñez, J.A.; Fu, X.; Gu, P.; Fu, B. Exploring transfer learning for deep learning polyp detection in colonoscopy images using YOLOv8. *Comput. Aided Diagn.* **2025**, *13407*, 444–453.
- Li, W.; Zhao, W.; Du, Y. Large-scale wind turbine blade operational condition monitoring based on UAV and improved YOLOv5 deep learning model. *Mech. Syst. Signal Process.* **2025**, *226*, 112386. [\[CrossRef\]](#)
- Wang, C.Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
- Wen, M.; Li, C.; Xue, Y.; Xu, M.; Xi, Z.; Qiu, W. YOFIR: High precise infrared object detection algorithm based on YOLO and FasterNet. *Infrared Phys. Technol.* **2025**, *144*, 105627. [\[CrossRef\]](#)
- Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R. BiFormer: Vision Transformer with Bi-Level Routing Attention. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
- Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
- Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [\[CrossRef\]](#)
- Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *arXiv* **2017**, arXiv:1707.01083. [\[CrossRef\]](#)
- Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020.
- Chen, J.; Kao, S.-h.; He, H.; Zhou, W.; Wen, S.; Lee, C.-H. Run, Don’t Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
- Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [\[CrossRef\]](#)
- Cheng, M.; Cai, K.; Li, M. RWF-2000: An Open Large Scale Video Database for Violence Detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021.

22. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
23. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Mallocci, M.; Kolesnikov, A.; et al. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.* **2018**, *128*, 1956–1981. [[CrossRef](#)]
24. Goodfellow, I.J.; Bulatov, Y.; Ibarz, J.; Arnoud, S.; Shet, V. Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. *arXiv* **2013**, arXiv:1312.6082.
25. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the Computer Vision ECCV 2018: 15th European Conference, Munich, Germany, 8–14 September 2018; Part XIV. Volume 11218, pp. 122–138.
26. Chen, H.; Wang, Y.; Guo, J.; Tao, D. VanillaNet: The Power of Minimalism in Deep Learning. *arXiv* **2023**, arXiv:2305.12972. [[CrossRef](#)]
27. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 10096–10106.
28. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; Part VII. Volume 11211, pp. 3–19.
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
31. Rezatofighi, H.; Tsoi, N.; Gwak, J.Y.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019.
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *arXiv* **2019**, arXiv:1911.08287. [[CrossRef](#)]

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.