information

Article

S-YOLOv5: A Lightweight Model for Detecting Objects Thrown from Tall Buildings in Communities

Yuntao Shi [1,2], Qi Luo [1,2], Meng Zhou [1,2], Wei Guo [1,2,*], Jie Li [1,2], Shuqin Li [1,2] and Yu Ding [1,2]

[1] School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; shiyuntao@ncut.edu.cn (Y.S.); laki@mail.ncut.edu.cn (Q.L.); zhoumeng@ncut.edu.cn (M.Z.); lijie1986@ncut.edu.cn (J.L.); lsq@ncut.edu.cn (S.L.); dingyu@ncut.edu.cn (Y.D.)
[2] Key Lab of Field Bus and Automation of Beijing, North China University of Technology, Beijing 100144, China
* Correspondence: guowei0903@ncut.edu.cn

Abstract: Objects thrown from tall buildings in communities are characterized by their small size, inconspicuous features, and high speed. Existing algorithms for detecting such objects face challenges, including excessive parameters, overly complex models that are difficult to implement, and insufficient detection accuracy. This study proposes a lightweight detection model for objects thrown from tall buildings in communities, named S-YOLOv5, to address these issues. The model is based on the YOLOv5 algorithm, and a lightweight convolutional neural network, Enhanced ShuffleNet (ESNet), is chosen as the backbone network to extract image features. On this basis, the initial stage of the backbone network is enhanced and the simplified attention module (SimAM) attention mechanism is added to utilize the rich position information and contour information in the shallow feature map to improve the detection of small targets. For feature fusion, the sparsely connected Path Aggregation Network (SCPANet) module is designed to use sparsely connected convolution (SCConv) instead of the regular convolution of the Path Aggregation Network (PANet) to fuse features efficiently. In addition, the model uses the normalized Wasserstein distance (NWD) loss function to reduce the sensitivity of positional bias. The accuracy of the model is further improved. Test results from the self-built objects thrown from tall buildings dataset show that S-YOLOv5 can detect objects thrown from tall buildings quickly and accurately, with an accuracy of 90.2% and a detection rate of 34.1 Fps/s. Compared with the original YOLOv5 model, the parameters are reduced by 87.3%, and the accuracy and rate are improved by 0.8% and 63%, respectively.

Keywords: lightweight; objects thrown from tall buildings; sparsely connected convolution; attention mechanism

Citation: Shi, Y.; Luo, Q.; Zhou, M.; Guo, W.; Li, J.; Li, S.; Ding, Y. S-YOLOv5: A Lightweight Model for Detecting Objects Thrown from Tall Buildings in Communities. *Information* **2024**, *15*, 188. https://doi.org/10.3390/info15040188

Academic Editor: Danilo Avola

Received: 19 January 2024
Revised: 21 February 2024
Accepted: 4 March 2024
Published: 29 March 2024

1. Introduction

The detection of objects thrown from tall buildings in communities is essential for public safety and social governance. With the acceleration of urbanization and the improvement of urban infrastructure, the number of high-rise residences is increasing. However, accidents involving objects thrown from a height have also increased, becoming a severe problem in contemporary urban development. Objects thrown from tall buildings endanger both personal safety and property. Therefore, it is crucial to trace the source of incidents of objects being thrown from tall buildings. Currently, the leading solution is to install video surveillance equipment on the external walls of high-rise residential buildings to collect evidence and deter potential objects being thrown from tall buildings. However, it is difficult to identify the behavior of objects thrown from tall buildings quickly and accurately by only relying on traditional video surveillance technology [1]. In the field of the detection of objects thrown from tall buildings, visual images provided by visible light signals have a high resolution, detail-rich information, and high real-time performance, so visual images have significant advantages in dealing with the problem of objects thrown from tall buildings. In recent years, target detection technology based on visual images has

aroused extensive research interest, providing a new solution to this social problem. Using computer vision technology to identify objects thrown from tall buildings in real time can effectively improve the monitoring system's performance and achieve the rapid detection of objects thrown from tall buildings.

Communities' demand for the lightweight and efficient detection of objects thrown from tall buildings is currently difficult to meet, and existing lightweight detection models [2] perform poorly in dealing with the problems of the small size of objects, inconspicuous features, and high speed. To address this, this paper proposes an improved YOLOv5 detection model specifically for the complex detection of objects thrown from tall buildings, and its main contributions include the following:

(1) In this paper, we design the S-ESNet backbone network to improve the performance of small target detection. The network accomplishes this by enhancing initial features and integrating the SimAM attention mechanism.
(2) An SCPANet module is proposed, utilizing SCConv architecture to achieve efficient object detection on high-resolution feature maps.
(3) To reduce the sensitivity of the Intersection over Union (IoU) of objects thrown from tall buildings to target position deviation, the NWD loss function is introduced to improve the model's accuracy.

2. Related Works

Deep learning has received significant attention in image processing and target detection in recent years due to its superior feature extraction capabilities. While developing this field, researchers have proposed numerous deep-learning-driven target detection methods dedicated to achieving a balance between lightweighting the model, increasing the computational speed, reducing the number of parameters, and improving the target detection accuracy. In 2016, Iandola et al. [3] proposed a SqueezeNet model that used small kernel convolution to compress the feature dimensions and verified the model's effectiveness on the ImageNet dataset, significantly reducing the number of parameters and maintaining the detection accuracy. From 2017 to 2019, Zhu et al. [4–6] developed the MobileNet family of models using deeply separable convolution and Squeeze-and-Excitation (SE) attention mechanism [7]; they tested the models on the ImageNet dataset and showed the efficiency and faster speed of the models. In 2018, Ma et al. [8] designed the ShuffleNet network, effectively reducing the computational effort of point-by-point convolution through group convolution and channel blending techniques. They verified the efficient performance of the network on the ImageNet dataset. In 2019, Tan et al. [9] proposed the MnasNet network, demonstrating its performance on ImageNet through multi-objective optimization and hierarchical search space optimization for excellent performance and flexibility. In 2020, Han et al. [10] proposed the GhostNet network, which reduces the computational effort in the ImageNet dataset by generating Ghost feature maps while demonstrating increased computational efficiency. In the same year, Tan et al. [11] proposed the EfficientDet model, which utilizes a Bidirectional Feature Pyramid Network (BiFPN) and composite scaling methods to achieve fewer model parameters and higher running speeds on the COCO dataset. In 2021, Zhu et al. [12] improved YOLOv5 by introducing Transformer Prediction Heads [13] and convolution block attention model (CBAM) [14] on the VisDrone2021 dataset, which improved the accuracy and recall of target detection. In 2022, Ma et al. [15] proposed MoCoViT network, which adopts the Mobile Self-Attention mechanism (MoSA) and Mobile Feed Forward Network (MoFFN), and through experiments on the COCO dataset, it is demonstrated that the network can maintain good performance while reducing model complexity and memory footprint.

The lightweight model studied above reduces the number of layers and channels in the high-resolution stage by rapidly reducing the sampling rate of the feature map. While this effectively reduces the amount of computation, it results in a significant loss of rich feature information, which reduces the accuracy of small target detection. The models optimized for small objects, on the other hand, despite improving the detection performance, also

subsequently increase the computational burden and limit their applicability in community applications for the detection of objects thrown from tall buildings due to the limitation of computational resources.

In this study, a lightweight detection model, S-YOLOv5, is proposed to address the challenge of the existing algorithms in detecting objects thrown from tall buildings, which makes it challenging to satisfy both detection accuracy and operation speed. The model employs the ESNet [16] backbone network to reduce the model size and enhance the detection accuracy by strengthening the initial phase of the backbone network in combination with the SimAM [17] attention mechanism. Meanwhile, lightweight optimization is performed at the feature fusion layer. In addition, the model integrates NWD loss [18] to reduce the sensitivity of IoU [19] to small target localization bias, further improving the accuracy. Through experimental validation, the model demonstrates high detection accuracy and fast processing capability on the self-constructed objects thrown from tall buildings dataset, which is suitable for the real-time monitoring of the behavior of objects thrown from tall buildings.

3. Lightweight Model S-YOLOV5

3.1. S-YOLOV5 Network Architecture

Given the small size of objects thrown from tall building detection targets, insufficient feature information, and fast speed change, the existing detection scheme lacks target recognition performance. It cannot meet communities' requirements for real-time detection. For this reason, this paper proposes a novel object thrown from tall buildings detection model, S-YOLOv5 (Figure 1), based on the YOLOv5 network architecture, aiming to accelerate the detection speed and improve the accuracy. The model consists of four parts: input layer, backbone network, neck network, and head. The backbone network is responsible for extracting features, while the neck network is responsible for the fusion of features.

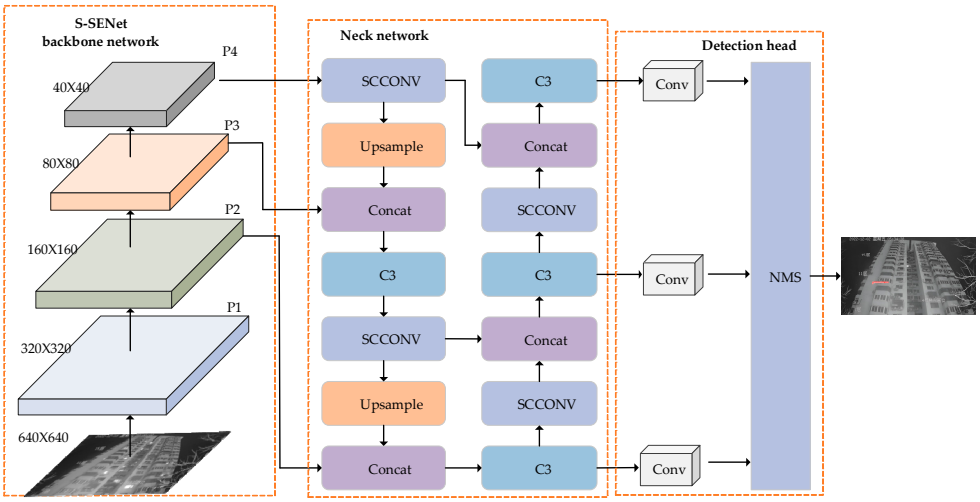


Figure 1. S-YOLOv5 network structure.

In this paper, improvements are made to the backbone network and the neck network. For the backbone network, two extra ES Block modules are added after the first downsampling, and the SE attention mechanism is optimized to the SimAM attention mechanism to construct a novel S-SENet backbone network. Compared with the original architecture of YOLOv5, S-SENet improves the detection accuracy of tiny objects thrown from tall buildings while keeping the model lightweight. The PANet [20] structure is improved for the neck network by replacing the traditional convolution with SCConv [21]. This adjustment slightly reduces the detection accuracy, reduces the number of model parameters, and improves the detection efficiency. In addition, NWD Loss is introduced to

reduce the sensitivity of the model to the positional deviation of objects thrown from tall buildings, which further enhances the detection accuracy of the model.

3.2. Backbone Network S-ESNet

Initial high-resolution features carry important detail information in the backbone network, which is critical for the identification and localization of small targets. Current lightweight backbone networks tend to rapidly reduce the resolution of these features, resulting in the loss of many layers and channels while maintaining a high resolution. Although this reduces the computational complexity, it also means that a lot of feature information is lost. To improve the detection performance of small targets, this study proposes a novel lightweight backbone network called S-ESNet, the structure of which is illustrated in Figure 2. The network mainly consists of one (Convolution, Batch Normalization, and Hardswish Activation) CHB module, three SES2_1 modules, and four SES1_X modules, where X denotes 1, 2, or 5. Compared with the initial ESNet, S-ESNet adds two extra ES Block modules in the first downsampling stage. By delaying the primary stage of downsampling and increasing the computational resources, the network can extract and save more low-level semantic information [22], which in turn improves the detection accuracy of objects thrown from tall buildings.

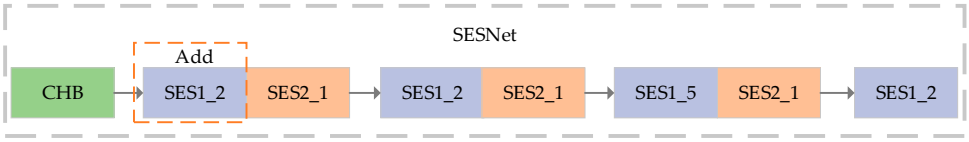


Figure 2. S-ESNet backbone network.

The backbone network's core improvement optimizes the attention mechanism in the ES Block structure. Like MobileNetV3, ESNet also applies the SE attention mechanism in all blocks, strengthening the ability to characterize target-related features through adaptive weighting of network channels. However, the SE attention mechanism is weak when dealing with spatial dimensions, which limits the model's performance in the detection of objects thrown from tall buildings.

To further improve the performance of ESNet in this application scenario, this study improves the attention mechanism in the ES Block from SE to SimAM based on the feature extraction network architecture of ESNet. As shown in Figure 3, the improved part of the ES Block is that the SimAM attention mechanism generates a unified 3D attention weight simultaneously, which enables the model to focus more on deeper information about objects thrown from tall buildings in the feature space. SimAM is based on the phenomenon of spatial inhibition in neuroscience, which measures the information density of a neuron and assigns weights by minimizing an energy function, where the minimum energy function of E, the nth neuron, can be formulated as

$$\begin{cases} E = -\frac{4\left(\hat{\sigma}^2+\lambda\right)}{(t-\hat{\mu})^2+2\hat{\sigma}^2+2\lambda} \\ \hat{\sigma}^2 = 1/M\sum_{i=1}^{M}(x_i-\hat{\mu})^2 \\ \hat{\mu} = 1/M\sum_{i=1}^{M}x_i \end{cases} \tag{1}$$

where $\hat{\sigma}^2$ represents the variance of all neurons, λ is the regularization coefficient, t is the target neuron, $\hat{\mu}$ is the search distance, M denotes the number of neurons on each channel, and x is the neighboring neurons.

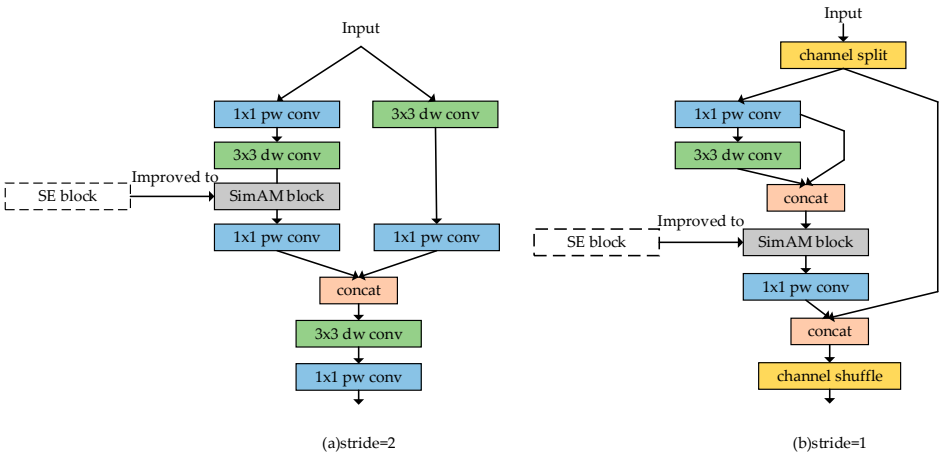


Figure 3. Improved ES Block.

Equation (1) shows that when the energy value is low, the more significant the difference between neuron x and its surrounding neurons, the higher the importance. The higher the weight value assigned to x, the greater the value of E is. According to the principle of the attention mechanism, the Sigmoid function is used to normalize the pair, i.e., the weight of each neuron can be expressed as sigmoid$(1/E)$. Finally, the input feature layer performs a Hadamard product operation with attention weights to obtain the augmented attention feature map.

By replacing the attention mechanism in the ES Block from SE to SimAM, this study comprehensively evaluates spatial location features and channel information without increasing the parameters and effectively focuses on essential neurons. This improvement helps to enhance the feature extraction efficiency and training speed of the network.

3.3. Feature Fusion Module

The feature enhancement module PANet employs a bidirectional multilevel fusion strategy to integrate bottom-up and top-down features. Given the computational enhancement of PANet due to the backbone network enhancement, this paper proposes SCConv to reduce the computational burden, which consists of a combination of two different types of convolutional structures, as demonstrated in Figure 4. The first one is channel-by-channel convolution (DWConv), which applies a different convolution kernel to each input channel separately, achieves the convolution operation of a single convolution kernel with a single channel, and subsequently integrates the output results by point-by-point convolution. The second is Pointwise Grouped Convolution (PWGConv) [23], which has similar functionality to standard 1×1 convolution.

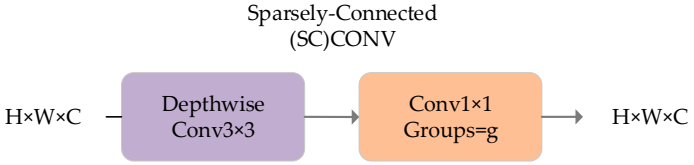


Figure 4. SCConv network structure.

The comparison of the number of convolution parameters before and after the improvement can be formulated as

$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F / g}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K^2 g} \tag{2}$$

where M represents the number of input channels, N represents the number of output channels, D_F represents the size of the input feature map, D_K represents the size of the

convolutional kernel. The molecule in the formula represents the sum of the channel-by-channel convolution computation amount and the point-by-point convolution computation amount, and the denominator represents the standard convolution computation amount. The number of convolutional kernels, N, is generally large during network training, so the proportion of parameters in the formula is mainly affected by the convolutional kernel D_K and the number of groupings of grouping convolutions. Therefore, the parameter amount of deep separable convolution in the SCConv module is about $1/D_K^2 g$ of the ordinary convolution. In summary, SCConv can effectively reduce the connection between channels, effectively reduce the parameter amount of the model, and improve the model's performance in terms of the detection speed.

In this study, a novel lightweight feature enhancement module, named SCPANet, is constructed based on SCConv, as demonstrated in Figure 5. In the PANet structure, SCPANet reduces the computational overhead by replacing the traditional 3×3 convolutional layers.

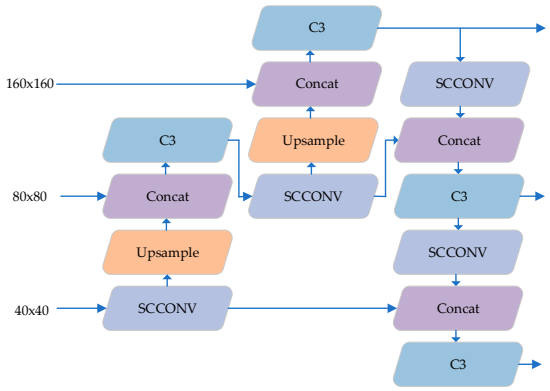
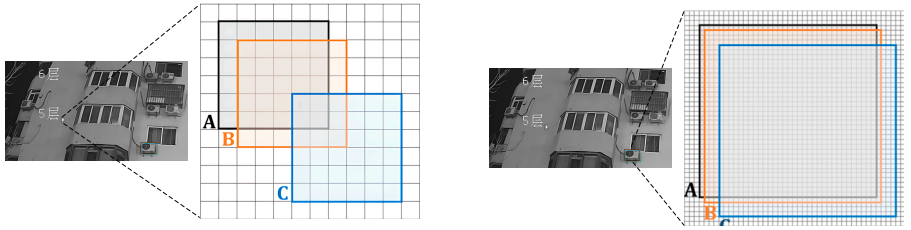


Figure 5. SCPANet network structure.

3.4. NWD Loss Function

Figure 6 demonstrates that objects thrown from tall buildings are typically small and consist of limited pixels, which results in significant IoU fluctuations—ranging from 0.53 to 0.06—even with minor positional changes. Such fluctuations markedly impact the accuracy of label assignments. By contrast, larger objects that contain more pixels experience less variation in IoU; a similar positional offset might only reduce the IoU from 0.90 to 0.65. The conventional IoU methods and their variations thus show a high susceptibility to the position deviation of these small targets, potentially causing a noticeable drop in the performance of anchor-based detectors.



$$IoU = \frac{|A \cap B|}{|A \cup B|} = 0.53 \Rightarrow IoU = \frac{|A \cap C|}{|A \cup C|} = 0.06 \qquad IoU = \frac{|A \cap B|}{|A \cup B|} = 0.90 \Rightarrow IoU = \frac{|A \cap C|}{|A \cup C|} = 0.65$$

Figure 6. The sensitivity analysis of IoU on tiny and normal scale objects.

To solve the above problem, this paper introduces the NWD loss function, which models the bounding box as a two-dimensional Gaussian distribution and evaluates the similarity between the bounding box and the actual bounding box by predicting their

corresponding Gaussian distributions. The distribution similarity can be utilized to evaluate whether the detected targets overlap or not. The normalized Wasserstein distance can be formulated as

$$\mathrm{NWD}(N_A, N_B) = \exp\left(-\frac{\sqrt{w_2^2(N_A, N_B)}}{C}\right) \tag{3}$$

where N_A and N_B are Gaussian distributions modeled by $A = (cx_A, cy_A, w_A, h_A)$ and $B = (cx_B, cy_B, w_A, h_A)$, $W_2^2(N_A, N_B)$ is the distance measure, C is a constant closely related to the dataset. Since NWD is insensitive to the scale of the objects, it has an outstanding advantage in measuring the similarity between small targets thrown from tall buildings. The NWD loss is introduced into the regression loss function to make up for the deficiency of the IoU loss in the detection process of small targets, and at the same time, the ratio of the IoU to the NWD loss is adjusted to 8:2, as shown in Equation (4). The above improvements to the loss function help improve the model's detection performance for small objects thrown from tall buildings.

$$Loss_{\mathrm{loc}} = IoU_{\mathrm{loss}} \times 0.8 + NWD_{\mathrm{loss}} \times 0.2 \tag{4}$$

4. Experimental Results and Analysis

4.1. Dataset Construction

This paper collects a dataset of objects thrown from tall buildings using community surveillance and selfie images. It encompasses six types of objects: cigarette butts, stones, plastic bottles, cans, knives, and clothing. The dataset captures a wide range of community scenes, various viewpoints, different weather conditions, and multiple throwing actions, enhancing the model's ability to generalize. Many data augmentation techniques—such as level flipping, Gaussian blur, random translation, and affine transformation—are employed to combat model overfitting during training. These techniques are randomly combined to enrich the dataset. Figure 7 presents examples of the thrown object images. The dataset includes 11,064 images, segmented into a training set with 6638 images, a validation set with 2213 images, and a test set comprising 2213 images, following a 6:2:2 ratio. The image count for each object type is 2561 for cigarette butts, 1047 for stones, 2410 for plastic bottles, 2310 for cans, 1230 for knives, and 1506 for clothing.

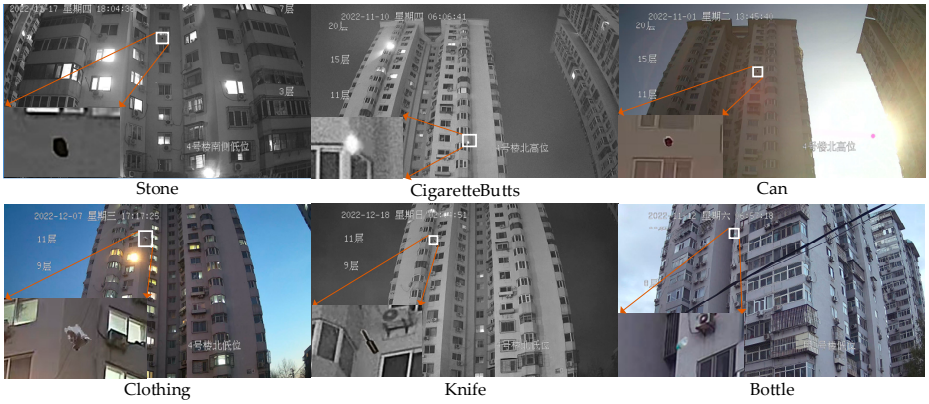


Figure 7. Partial images in the dataset.

4.2. Experimental Environment

In this experiment, Python is selected as the programming language, and the detection model is built using the PyTorch [24] deep learning library. The operating system deployed for conducting the experiments is CentOS 7.9, equipped with an Intel Xeon Silver 4210R processor and an NVIDIA Quadro RTX 5000 GPU. Detailed configurations of the relevant experimental parameters are provided in Table 1.

Table 1. Experimental parameter settings.

Parameter	Parameter Value
Batch size	8
Image size	640 × 640
Learning rate	0.01
Momentum	0.935
Attenuation coefficient	0.0004
Iteration rounds	100

4.3. Evaluating Indicator

In this experiment, the Mean Average Precision (*mAP*) is used as a criterion for evaluating the recognition accuracy of the model. The mAP is closely related to the accuracy (*P*) and recall (*R*) of the model [25], which is calculated as follows:

$$P = \frac{T_P}{T_P + F_P} \tag{5}$$

$$R = \frac{T_P}{T_P + F_N} \tag{6}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{7}$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \times 100\% \tag{8}$$

where T_P is the number of positive samples predicted correctly, F_P is the number of positive samples mispredicted, and F_N is the number of negative samples mispredicted. *AP* is the integral of the P-R curve, and the area under the curve is the *AP* value. The *mAP* is the metric obtained by averaging the average precision (*AP*) [26] across all categories; *C* represents the number of categories.

To compare the real-time detection speeds of different models, this study adopts the number of transmitted frames per second (FPS) as the performance evaluation index. With the FPS metric, the detection speed of the models can be visualized, and then the real-time detection performance of the models can be evaluated more accurately.

4.4. Model Training

YOLOv5 utilizes a K-means clustering algorithm to generate anchor frames automatically based on a statistical analysis of different-sized targets in the training set. This study uses stochastic gradient descent (SGD) as the optimizer and employs a cosine annealing strategy to regulate the learning rate [27]. The same parametric strategy is adopted to train the S-YOLOv5 model. As shown in Figure 8, the model performs better in training and validation loss than the original YOLOv5, which reduces the training loss and validation loss by 0.0062 and 0.0020, respectively.

By analyzing the accuracy curves at different scales, as shown in Figure 9, the S-YOLOv5 model outperforms the original model in terms of accuracy at multiple scales. With the improvement, P and R are improved by 1.4% and 0.86%, respectively. When the IOU threshold is set to 0.5, mAP0.5 is improved by 0.8%. At different IOU threshold ranges (from 0.5 to 0.9), mAP0.5:0.9 increased by 0.5%. The results demonstrate that the proposed lightweight improvement measures effectively enhance the detection performance of the YOLOv5 network, enhancing the model's accuracy in detecting objects thrown from tall buildings.

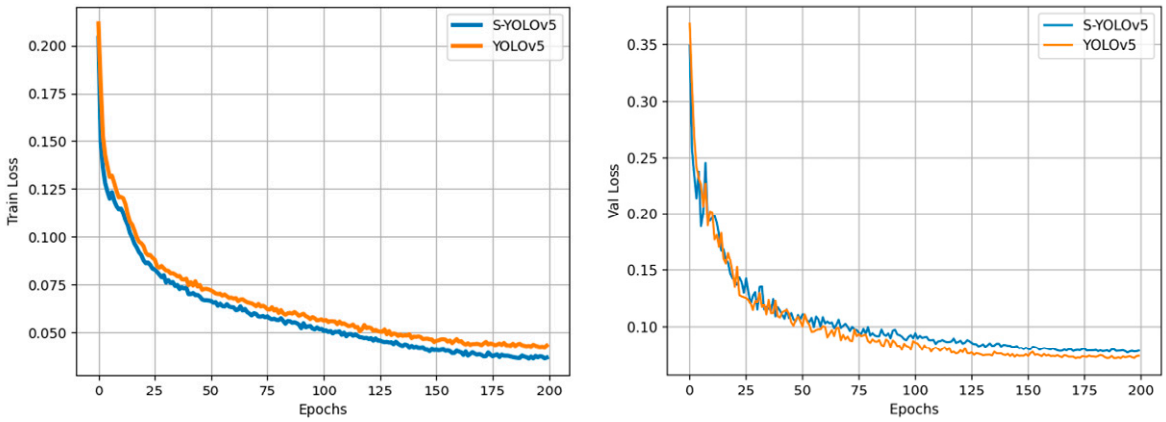


Figure 8. The curve of model loss.

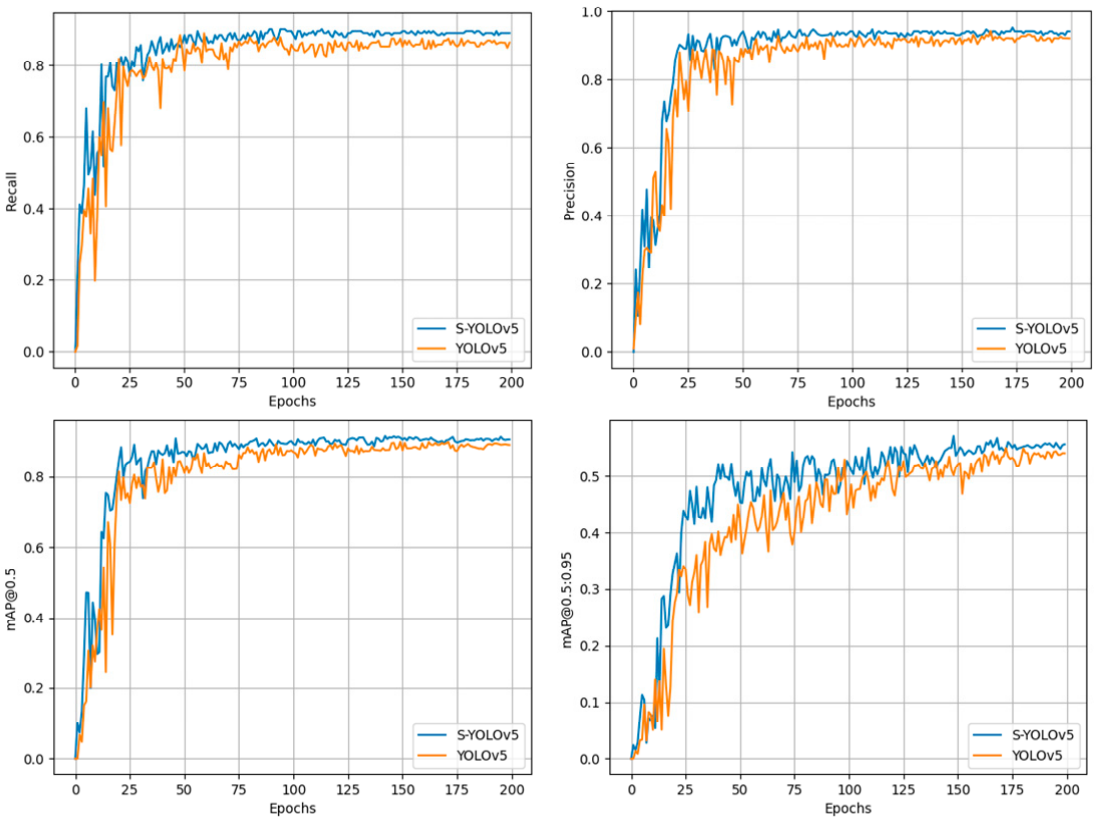


Figure 9. The curve of model accuracy.

The confusion matrix is a two-dimensional matrix that expresses the relationship between the model's predicted categories and the actual label categories. The confusion matrix shown in Figure 10 indicates that clothing detection achieved the highest accuracy of 93%. Cigarette butts, stones, plastic bottles, cans, and knives were also all detected with over 85% accuracy. The overall performance is relatively balanced. However, the fact that most objects thrown from tall buildings are small and poorly characterized and are often captured under conditions with complex backgrounds in natural environments results in some objects thrown from tall buildings being misidentified.

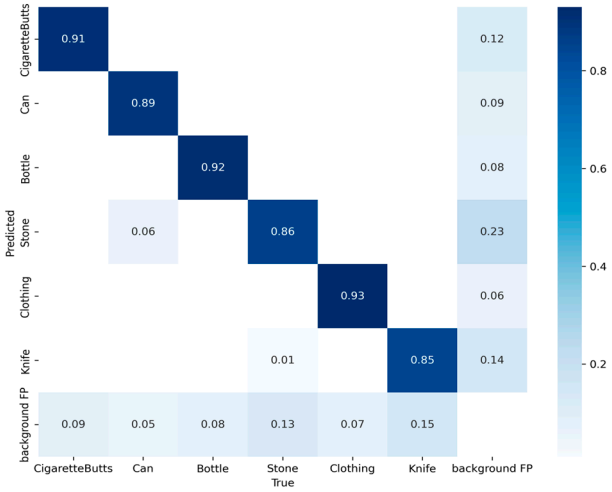


Figure 10. Confusion matrix of S-YOLOv5.

This study compares the performance of the improved S-YOLOv5s model with the initial YOLOv5s model by evaluating it on the validation and test sets containing scenarios of objects being thrown from tall buildings. According to Table 2, the S-YOLOv5s model outperforms the original YOLOv5s in terms of precision (P), recall (R), and average precision (AP). S-YOLOv5s improves P by 1.12% and 1.09%, R by 0.92% and 0.91%, and AP by 0.94% and 0.92% on both the validation and the test sets. The experimental data show that the improved algorithm enhances the model's ability to accurately localize and recognize objects thrown from tall buildings with improved accuracy.

Table 2. Performance comparison between S-YOLOv5 and YOLOv5 on test and validation sets.

Dataset	Model	P/%	R/%	mAP/%
val	YOLOv5	89.12	86.61	89.25
	S-YOLOV5	90.24	87.53	90.19
test	YOLOv5	89.14	86.75	89.33
	S-YOLOV5	90.23	87.66	90.25

4.5. Comparison Experiment

To validate the effectiveness of the proposed detection model in this study, a comparative analysis with five other detection models (Faster R-CNN [28], Efficientdet [29], and YOLOv7-tiny [30], YOLOv5s) is conducted. All models use the same dataset of objects thrown from tall buildings with the same training and validation strategy, and the experimental results are shown in Table 3.

Table 3. Comparative experiment.

Model	Backbone Network	mAP/%	Parameters/M	FPS/S
Faster R-CNN	ResNet50	89.9	108.9	2.7
SDD	Vgg16	83.6	100.3	20.4
Efficientdet	EfficientNet-B0	86.7	15.1	13.5
YOLOv7-tiny	Darknet	89.1	6.2	22.8
YOLOv5s	C3	89.4	13.7	20.9
S-YOLOv5	S-ESNet	90.2	1.74	34.1

Table 3 shows that the S-YOLOv5 model outperforms Faster R-CNN, SSD, EfficientDet, YOLOv5, and YOLOv7-tiny regarding the mAP, number of parameters, and detection speed. Compared to these models, S-YOLOv5 improves on AP by 0.3%, 6.6%, 3.5%, 1.1%, and

0.8%. Although Faster R-CNN performs well in terms of accuracy as a two-stage model, its large number of parameters results in a low detection speed of 2.7 Fps/s, which is unsuitable for community security applications requiring real-time detection. While the single-stage model improves the speed, it cannot match Faster R-CNN in terms of accuracy. In contrast, the optimized S-YOLOv5 model increases the detection speed to 34.1 Fps/s while maintaining a high accuracy of 90.2%.

This study's proposed S-YOLOv5 model demonstrates excellent performance in standard reviews. The model is based on the YOLOv5 algorithm, which employs a lightweight ESNet convolutional neural network as the core network to enhance the feature extraction efficacy. S-YOLOv5 enhances the initial stage of the core network and incorporates the SimAM attention mechanism, which enables the model to mine the rich location and contour information of the shallow feature maps more efficiently, and therefore improves the detection of small targets with better accuracy. In the feature integration module, this paper achieves the efficient integration of features by cleverly designing SCPANet and using SCConv. This innovative strategy can effectively improve the target detection efficiency. Meanwhile, a novel NWD loss function is also incorporated into the study to minimize the impact on the localization error, improving the detection accuracy. The results of the model detecting objects thrown from tall buildings are shown in Figure 11, which show it can effectively recognize cigarette butts, stones, plastic bottles, cans, knives, and clothes. In conclusion, the S-YOLOv5 model proposed in this paper can accurately and quickly recognize objects thrown from tall buildings and meet the real-time detection requirements.

Figure 11. Detection results of S-YOLOv5.

4.6. Ablation Experiment

To verify the performance improvement caused by the improvements in the S-YOLOv5 model to the original network model, three improvement methods, namely, S-ESNet backbone network, SCPANet module, and NWD loss function, are sequentially added to the original network model. Ablation experiments are carried out with the same dataset and training strategy, and the results of the experiments are shown in Table 4.

Table 4. Comparative experiment.

Model	S-ESNet	NWD Loss	SCPANet	mAP/%	Parameter/M	FPS/S
Original model				89.4	13.7	20.9
Exp1	√			88.8	2.08	32.6
Exp 2	√	√		90.5	-	30.6
Exp3	√	√	√	90.2	1.74	34.1

Experiment I uses the S-ESNet network to lighten and improve the original backbone network by enhancing its early features and introducing SimAM. Although the replacement model has a 0.6% decrease in the average accuracy relative to the YOLOv5 model, the

parameters are reduced by 84.8%, and the detection speed is improved by 55.9%. This indicates that Experiment I successfully improved the running speed of the model at the expense of a small amount of accuracy. Experiment II introduces NWD loss on top of Experiment I. Although the model's speed decreases slightly, the average accuracy improves by 1.7% compared to Experiment I. This is mainly because NWD loss helps improve the model's accuracy. This is mainly because NWD loss helps to mitigate the sensitivity to the deviation in the position of objects thrown from tall buildings during detection. Experiment III introduces the SCPANet necking network based on Experiment II to improve the real-time detection speed of the model, which improves the detection speed by 11.4% compared to Experiment II. However, a small amount of accuracy needs to be sacrificed. This paper explores three step-by-step improvement strategies to achieve better target detection performance by maintaining high accuracy while lightweighting the model.

5. Conclusions

In this paper, a new lightweight detection algorithm for objects thrown from tall buildings named S-YOLOv5 is proposed by systematically analyzing the influencing factors of lightweight algorithms on the accuracy of the detection of objects thrown from tall buildings. To cope with the problem of the slow running speed of the detection model and the difficulty of adapting to lightweight deployment in communities, this paper designs a new backbone network, S-ESNet. It introduces the SCPANet neck network, which significantly improves the detection speed. This paper adopts several strategies to address the problem that objects thrown from tall buildings are small and difficult to detect. In the first strategy, the shallow feature information of objects thrown from tall buildings is preserved by enhancing the early features of the backbone network and introducing SimAM. Secondly, NWD loss is introduced to effectively reduce the sensitivity to the positional deviation of the IoU of the detection targets of objects thrown from tall buildings, improving the model's convergence speed and detection accuracy. The experimental results show that the S-YOLOv5 algorithm is better than YOLOv5 in detecting objects thrown from tall buildings. At the same time, more accurate feature recognition and detection performance can be achieved for small objects thrown from tall buildings. However, there are some shortcomings in this study. Due to the single relatively limited dataset sample, it may be complex for the algorithm to accurately capture the exact data of small targets in the images of objects thrown from tall buildings, resulting in a possible reduction in the model detection accuracy. To improve model performance, future research needs to focus on accumulating more diverse and challenging non-cooperative sample data and redesigning the network structure to reduce the over-reliance on data size.

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

Funding: This work is sponsored by the National Key R&D Program of China [2023YFC3306400], the National Natural Science Foundation of China (62273007).

Data Availability Statement: The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration (from you or one of your Contributing Authors) by another publisher.

Conflicts of Interest: All authors declare that they have no conflicts of interest.

Abbreviations

Definition	Abbreviation
ESNet	Enhanced ShuffleNet
SimAM	simplified attention module
SCPANet	sparsely connected Path Aggregation Network
SCConv	sparsely connected convolution
PANet	Path Aggregation Network
NWD	normalized Wasserstein distance
IoU	Intersection over Union
SE	Squeeze-and-Excitation
BiFPN	help of Bidirectional Feature Pyramid Network
CBAM	convolution block attention model
MoSA	Mobile Self-Attention
MoFFN	Mobile Feed Forward Network
CHB	Convolution, Batch Normalization, and Hardswish Activation
DWConv	channel-by-channel convolution
PWGConv	point-by-point grouping convolution
P	precision
R	recall
AP	average precision
mAP	Mean Average Precision
SGD	stochastic gradient descent
FPS	frames per second

References

1. Zaman, A.; Huang, Z.; Li, W.; Qin, H.; Kang, D.; Liu, X. Artificial intelligence-aided grade crossing safety violation detection methodology and a case study in new jersey. *Transp. Res. Rec. J. Transp. Res. Board* **2023**, *2677*, 688–706. [CrossRef]
2. Azimjonov, J.; Kim, T. Stochastic gradient descent classifier-based lightweight intrusion detection systems using the efficient feature subsets of datasets. *Expert Syst. Appl.* **2024**, *237*, 121493. [CrossRef]
3. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
4. 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.
5. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
6. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V. Searching for mobilenetv3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 November 2019; pp. 1314–1324.
7. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
8. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
9. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q.V. MnasNet: Platform-aware neural architecture search for mobile. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2815–2823.
10. 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; pp. 1577–1586.
11. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
12. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 10–17 October 2021; pp. 2778–2788.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; 2017; Volume 30. Available online: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 8 January 2024).

14. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
15. Ma, H.; Xia, X.; Wang, X.; Xiao, X.; Li, J.; Zheng, M. Mocovit: Mobile convolutional vision transformer. *arXiv* **2022**, arXiv:2205.12635.
16. Yu, G.; Chang, Q.; Lv, W.; Xu, C.; Cui, C.; Ji, W.; Dang, Q.; Deng, K.; Wang, G.; Du, Y.; et al. PP-PicoDet: A Better Real-Time Object Detector on Mobile Devices. *arXiv* **2021**, arXiv:2111.00902.
17. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 18–24 July 2021; pp. 11863–11874.
18. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized gaussian wasserstein distance for tiny object detection. *arXiv* **2022**, arXiv:2110.13389.
19. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-iou loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
20. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
21. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337. [CrossRef] [PubMed]
22. Zhang, Y.; Kang, W.; Liu, Y.; Zhu, P. Multi-scale semantic and detail extraction network for lightweight person re-identification. *Comput. Vis. Image Underst.* **2023**, *236*, 103813. [CrossRef]
23. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference On Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
24. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*; 2019; Volume 32. Available online: https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (accessed on 8 January 2024).
25. Zhu, C.; Liang, J.; Zhou, F. Transfer learning-based yolov3 model for road dense object detection. *Information* **2023**, *14*, 560. [CrossRef]
26. Bista, R.; Timilsina, A.; Manandhar, A.; Paudel, A.; Bajracharya, A.; Wagle, S.; Ferreira, J.C. Advancing tuberculosis detection in chest X-rays: A yolov7-based approach. *Information* **2023**, *14*, 655. [CrossRef]
27. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2017**, arXiv:1608.03983.
28. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
29. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single Shot Multibox Detector. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 22–37.
30. 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. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.