



Article

Long-Distance Person Detection Based on YOLOv7

Fan Tang ^{1,2,†} , Fang Yang ^{1,2,*,†}  and Xianqing Tian ^{1,2}¹ School of Cyberspace Security and Computer, Hebei University, Baoding 071000, China² Institute of Intelligence Image and Document Information Processing, Hebei University, Baoding 071000, China

* Correspondence: yangfang@hbu.edu.cn; Tel.: +86-1373-029-6453

† These authors contributed equally to this work.

Abstract: In the research field of small object detection, most object detectors have been successfully used for pedestrian detection, face recognition, lost and found, and automatic driving, among other applications, and have achieved good results. However, when general object detectors encounter challenging low-resolution images from the TinyPerson dataset, they will produce undesirable detection results because of the dense occlusion between people and different body poses. In order to solve these problems, this paper proposes a tiny object detection method TOD-YOLOv7 based on YOLOv7. First, this paper presents a reconstruction of the YOLOv7 network by adding a tiny object detection layer to enhance its detection ability. Then, we use the recursive gated convolution module to realize the interaction with the higher-order space to accelerate the model initialization process and reduce the reasoning time. Secondly, this paper proposes the integration of a coordinate attention mechanism into the YOLOv7 feature extraction network to strengthen the pedestrian object information and weaken the background information. Additionally, we leverage data augmentation techniques to improve the representation learning of the algorithm. The results show that compared with the baseline model YOLOv7, the detection accuracy of this model on the TinyPerson dataset is improved from 7.1% to 9.5%, and the detection speed reaches 208 frames per second (FPS). The algorithm of this paper is shown to achieve better detection results for tiny object detection.

Keywords: object detection; YOLOv7; recursive gated convolution; tiny object detection layer; coordinate attention mechanism



Citation: Tang, F.; Yang, F.; Tian, X. Long-Distance Person Detection Based on YOLOv7. *Electronics* **2023**, *12*, 1502. <https://doi.org/10.3390/electronics12061502>

Academic Editors: Mohamed Shehata and Mostafa Elhosseini

Received: 1 March 2023

Revised: 18 March 2023

Accepted: 21 March 2023

Published: 22 March 2023



Copyright: © 2023 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

Pedestrian detection is a crucial research area in computer vision, and thanks to advancements in deep learning technology, it has made unprecedented progress in recent years. Its application prospects have become more extensive, covering many fields, such as assistant driving, intelligent robots, intelligent transportation, rescue operations, and motion analysis [1,2]. However, the detection of tiny people objects (i.e., a single-person image less than 20 pixels [2,3]) requires thorough study, due to the significant challenges posed by dense occlusion between people and diverse poses [4]. There are two types of dense occlusion: inter-class occlusion, such as when people are blocked by cars, trees, or other objects, leaving only a tiny part of the human body exposed, and intra-class occlusion, where people walking on the street block each other. When people are reduced in proportion due to these types of occlusion, the background is enlarged, resulting in a scene with the dense occlusion of tiny people. In this case, the detection system may misjudge due to the huge and complex background, which can seriously impact the performance of the pedestrian detector [5]. Pedestrian detection generally requires a complete pedestrian object within the detection anchor. However, the human body is a non-rigid object that can adopt a range of postures, such as lying, sitting, and standing. This necessitates the network to identify the human body with different postures and provide an accurate full-body anchor; selecting the appropriate anchor is a significant research challenge.

Existing pedestrian detection datasets, such as CityPersons [6], mainly target short or medium distances, while INRIA primarily focuses on upright pedestrians' data, making them unsuitable for people for scenes with large areas and very long distances or for meeting the features of different pedestrian postures. The TinyPerson dataset aims at search and rescue scenes at sea or on the beach. The images in this dataset have the two significant difficulties mentioned above. Unlike objects of normal scale, tiny objects are more challenging to detect due to their small proportion in the image. Additionally, during the transmission process, the picture's or video's resolution may reduce and blur after the encoding and decoding process, causing the tiny objects to mix with the background. This requires a high computational power GPU (graphics processing unit) and much time to train [2], adding several challenges to producing qualified models.

This paper comprehensively considers the applicability and real-time performance of the object detection model in more scenarios and based on the features of quick and easy deployment of the one-stage algorithm YOLO (You Only Look Once) series [7] model reasoning. The speed and accuracy of YOLOv7 exceed all known object detectors from 5FPS to 160FPS [8]. This paper proposes an improved TOD-YOLOv7 model, which introduces the recursive gated convolution module to reduce the reasoning time by performing higher-order spatial interaction through gated convolution and recursive design [9]. Then, it reconstructs the network structure and adds a tiny object detection layer [10] to enhance the feature extraction network's ability to detect tiny objects. This paper further integrates a coordinate attention mechanism (CA) into the YOLOv7 feature extraction network to strengthen the information on pedestrian objects. This paper trains the TOD-YOLOv7 model from scratch using only the TinyPerson dataset and utilizes data augmentation methods, such as Mixup and Mosica, to improve the algorithm's representation learning. Compared with YOLOv7, this method achieves better object detection performance in ultra-low resolution images and outperforms it in detection tasks. The comparison of the effects is shown in Figure 1.

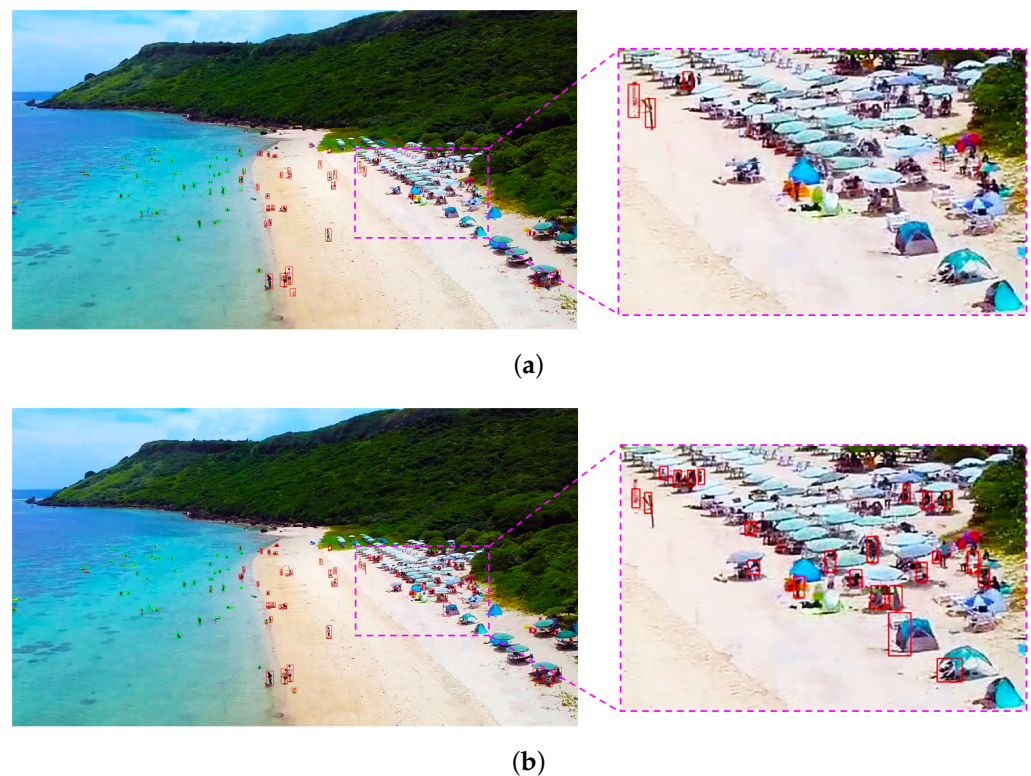


Figure 1. Intuitive examples on TinyPerson dataset [2] demonstrate the differences before and after the algorithmic improvements. (a) In the original YOLOv7 detection, many objects are missed in the enlarged detail. (b) This paper's improved algorithm TOD-YOLOv7 can detect more objects.

In summary, this article has the following contributions:

- This paper comprehensively analyzed the difficulties in the detection of tiny people, and proposed a network object detection scheme based on YOLOv7 to detect the objects in the TinyPerson image proposed in the context of rapid rescue at sea, providing a reference for new research.
- This paper reconstructed the YOLOv7 network by adding a tiny object detection head and combined it with recursive gating to explore its prediction potential with a self-attention mechanism. This reduces the reasoning time and improves the accuracy of the model. This paper also introduced a coordinate attention mechanism to address misdetections in tiny object detection, and combined the convolution attention mechanism and self-attention mechanism [11] to optimize the features.
- The proposed TOD-YOLOv7 improves the detection performance of the most advanced detector YOLOv7 by a significant improvement (2.4%).

2. Related Work

2.1. Dataset Description and Processing

Pedestrian detection is a popular algorithm in the field of computer vision and has widely been used [1]. Therefore, to achieve accurate and detailed results, it is necessary to have a pedestrian dataset with rich scenes, a large scale, and more detailed and accurate annotations for training and testing. Common datasets used for this purpose include MIT-CBCL, USC, Caltech-USA [12], DukeMTMC, INRIA [13], CityPersons [6] and Daimler [14]. The availability of these datasets promoted the development of pedestrian detection technology, and the continuous updating of these datasets reflects the desire for better datasets. However, many of these datasets are collected in urban scenarios and have high-resolution images with pedestrian objects occupying a large proportion, which is not conducive to training a model that can detect tiny objects. The TinyPerson dataset addresses this limitation, as it contains smaller-sized images with multiple poses and viewpoints, bringing greater complexity and making detection more difficult. Its most prominent feature is the presence of a deficient foreground-to-background object ratio with dense objects, with a person represented by a low-resolution object of fewer than 20 pixels in complex scenarios, such as a beach or the sea. Although this poses a challenge to the detection model, this diversity enables the model trained on the TinyPerson dataset to be well extended to more scenarios, leading to better performance in other systems.

Due to the small size of the object dataset used and the tiny size of the person objects being detected, to reduce the problem of too few samples and prevent over-fitting, this paper uses data augmentation. Currently, MixUp and Mosaic are considered effective data augmentation methods for image tasks in deep learning. An example of the data augmentation effect is shown in Figure 2. MixUp is carried out by interpolation, and the core idea is to randomly mix two training samples and their labels in a particular proportion.



Figure 2. Results of different data augmentation methods.

This hybrid method can increase sample diversity, smooth the transition of different types of decision boundaries, reduce the misrecognition of complex samples, improve

model robustness, and increase training stability. Mosaic combines multiple images into one image in a particular proportion. Similar to FMix [15], it belongs to the CutMix data augmentation algorithm. The resulting mosaic image has a higher level of detail and more labels, and training on it is equivalent to training on multiple tiny images, which allows the model to recognize objects in a smaller range and improve the detection performance of tiny objects.

2.2. Object Detection

The object detection algorithm typically samples numerous regions in the input image and then evaluates whether these regions contain objects of interest. The algorithm then adjusts the region boundary to more accurately predict the true boundary box of the object. At present, the commonly used object detection technology can be divided into two categories: anchor-based and anchor-free. The anchor-free algorithm directly predicts the location and size of the target through intensive prediction. However, this approach generates a large number of candidate boxes on the feature map, most of which are background boxes without the target. This increases the computational complexity and false detection rate. Moreover, small targets are challenging to detect because their size is very small. If the receptive field of the network is too large, it becomes difficult to obtain an accurate feature representation of small targets in the high-level feature map, which leads to a decrease in detection accuracy. The anchor-free algorithm is not suitable for detecting tiny objects in this paper, so the anchor-based algorithm is chosen. It can be further divided into two categories: one-stage and two-stage algorithms. The two-stage algorithm generates anchors with different sizes and proportions at each point of the feature map and then filters the anchors through a region proposal network (RPN) [16], such as Mask R-CNN [17], Faster R-CNN [16], Cascade R-CNN [18], etc. This approach provides high accuracy. On the other hand, the one-stage algorithm divides the original image into several grids and then obtains anchors of different sizes in each grid through a clustering method. It then determines the intersection over union (IOU) between the actual anchor and the predicted bounding box to obtain the training object, which is faster. Common one-stage algorithms include YOLO [19], YOLO9000 [7], YOLOV3 [20], YOLOV4 [21], YOLOV5, SSD [22], RetinaNet [23], etc. It is worth mentioning that although YOLOv5 has shown excellent performance in many scenarios, the YOLO official team has not published relevant papers on YOLOv5. Since its introduction in 2016, YOLO has widely been used in real-time systems for object recognition and positioning based on deep neural networks. With continuous iteration and improvement, the YOLO team has made significant strides in balancing speed and accuracy, making it the mainstream technology for object detection. YOLOv5 is considered a classic version, while YOLOv7 introduced model re-parameterization into the network architecture based on YOLOv5 and proposed a new efficient aggregation network architecture, ELAN (efficient long-range attention network), along with a training method that includes an auxiliary head [8,24]. This approach made YOLOv7 the most advanced object detector in the range of 5 FPS to 160 FPS. Therefore, this paper chose YOLOv7 as the baseline network in this research.

3. Tiny Person Detection Network

3.1. TOD-YOLOv7

The main network structure of TOD-YOLOv7 is depicted in Figure 3. The architecture consists of three parts: backbone, neck, and head. First, this paper continues to use the efficient aggregation network structure ELAN [8,24] in the backbone network to facilitate the network in learning more features and improving its robustness by controlling the shortest and longest gradient path. Building on this, this paper attempts to replace part of the ELAN structure with the recursive gated convolution ($g^n Conv$) module [9], which implements $g^n Conv$ based on convolution and avoids the secondary complexity of self-attention. The design of gradually increasing the channel width during the execution of spatial interaction also enables us to achieve high-order interaction with limited complexity.

Given that the foreground objects in TinyPerson data are typically ultra-low pixels, this study reconstructed the YOLOv7 network. Specifically, this paper incorporated a tiny object detection module [25,26] in the neck layers and added an extra tiny object detection header to improve the detection performance of the algorithm for tiny objects. Additionally, this paper integrated a coordinate attention module, which encodes the feature maps obtained from the upstream input separately and outputs them as a pair of directional perception and location-sensitive attention maps. These attention maps can be applied to the downstream input feature maps to enhance the network's ability to accurately locate and identify objects of interest.

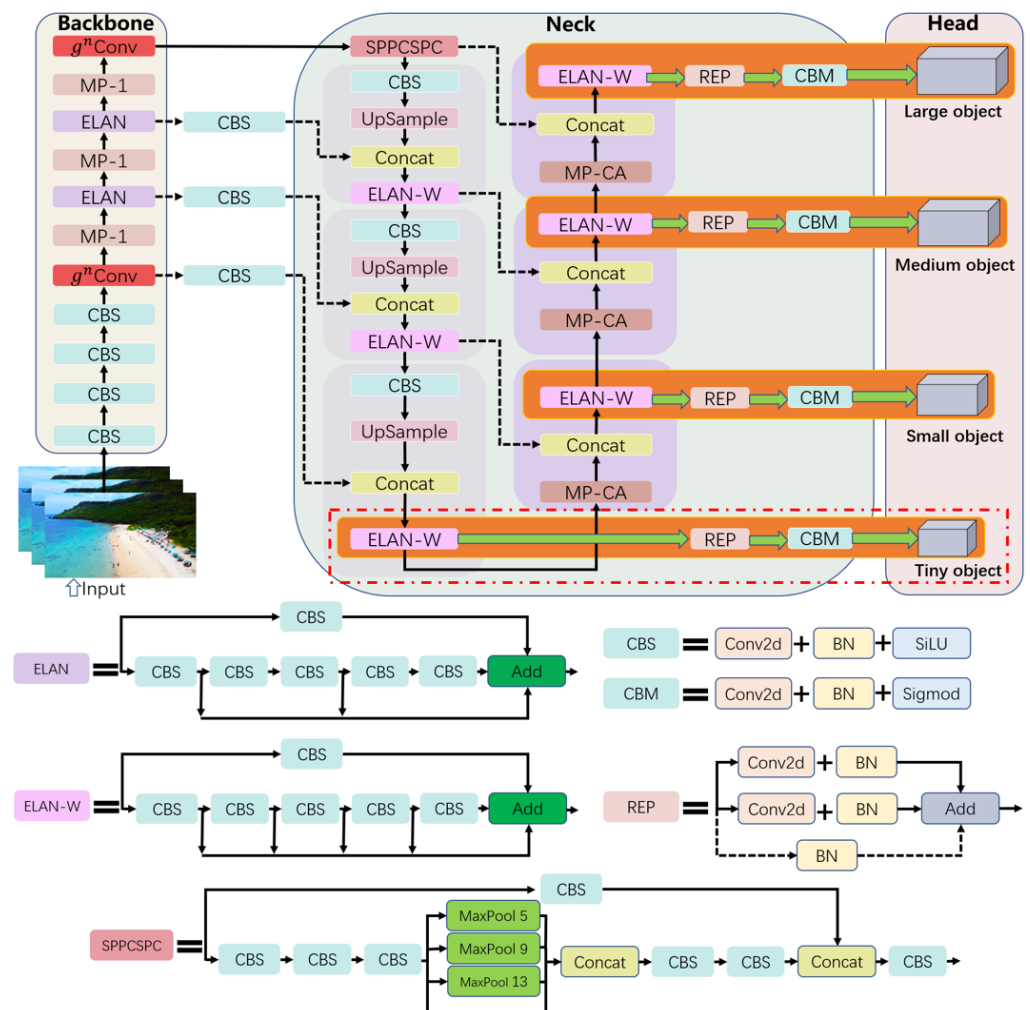


Figure 3. The network structure of TOD-YOLOv7. The red dotted box in the figure represents the extra detection head that this paper added to the model. The lower part of the figure illustrates the structural schematic diagram of specific components.

3.2. g^n Conv

In the field of tiny object detection, there are numerous instances where the size of the datasets needs to be increased to enhance the model's expression ability during training. This presents a challenge for the network model whose training dataset is not large enough. Among the features of a model, there can be complex and high-level interactions between any two spatial positions [9]. When this interaction is explicitly modeled in the model, it will improve the expression ability of the model. The success of self-attention [11] in visual transformer [27] proves this point. The idea of explicitly modeling higher-order spatial interaction is applied to CNN (convolutional neural network) so that neural networks can also complete higher-order spatial interaction. The recursive gated

convolution (g^nConv) module implements this concept by combining the recursive strategy and gated convolution ($gConv$); it is constructed using standard convolution through linear projection and element multiplication and performing spatial information interaction between the volume integration layer and full connectivity layers. This approach enhances the model's expression ability, which is crucial for successful detection in minuscule scenes. If this paper sets the upstream input feature as $x \in \mathbb{R}^{HW \times C}$, the output of the basic operation $gConv$ in g^nConv can be expressed as

$$\begin{aligned} [p_0^{HW \times C}, q_0^{HW \times C}] &= \phi_{in}(x) \in \mathbb{R}^{HW \times 2C}, \\ p_1 &= f(q_0) \odot p_0 \in \mathbb{R}^{HW \times C}, \\ y &= \phi_{out}(p_1) \in \mathbb{R}^{HW \times C}, \end{aligned} \quad (1)$$

In Equation (1), f is the depth convolution layer, ϕ_{in} and ϕ_{out} represent the linear projection's input and output operations, complete information interaction between adjacent features through element-wise multiplication, and output through linear projection. Here, this paper can express the first-order interaction between the feature $p_0^{(i)}$ (such as area A in Figure 4a) and its surrounding adjacent area $q_0^{(j)}$ (such as area B in Figure 4a) as

$$p_1^{(i,c)} = \sum_{j \in \Omega_i} w_{i \rightarrow j}^c q_0^{(j,c)} p_0^{(i,c)}, \quad (2)$$

where Ω_i is the current local window with i as the central coordinate, and w is the weight of depth convolution f . Each p_0 only interacts with adjacent feature q_0 once. In the same way, it is easy to achieve information interaction between long-distance and higher-order space and combine the features of each level in the adjacent areas around the object features to obtain $[q_0^{HW \times C_0}, \dots, q_{n-1}^{HW \times C_{n-1}}]$, it is can let the $\{q_k\}_{k=0}^{n-1} = [q_0^{HW \times C_0}, \dots, q_{n-1}^{HW \times C_{n-1}}]$, so

$$[p_0^{HW \times C_0}, q_0^{HW \times C_0}, \dots, q_{n-1}^{HW \times C_{n-1}}] = [p_0^{HW \times C_0}, \{q_k\}_{k=0}^{n-1}] = \phi_{in}(x) \in \mathbb{R}^{HW \times (C_0 + \sum_{0 \leq k \leq n-1} C_k)}. \quad (3)$$

Then let the gated convolution $gConv$ proceed recursively:

$$p_{k+1} = \frac{f_k(q_k) \odot g_k(p_k)}{\alpha}, k = 0, 1, \dots, n-1, \quad (4)$$

In Equation (4), f_k is a set of deep convolution layers, g_k is used to match the number of channels in each recursive process, and α is the scaling factor. In order to maintain the stability of training, divide the result by α to scale. In this paper, we give the calculation method of g_k :

$$g_k = \begin{cases} Identity, k = 0, \\ Linear(C_{k-1}, C_k), 1 \leq k \leq n-1. \end{cases} \quad (5)$$

When $k=0$, g_k is a certain value related to the specific network model. C_k sets the channel dimension of each order in the form of exponential decrement to reduce the excessive computational overhead in the process of higher-order information interaction. The calculation formula is as follows:

$$C_k = \frac{C}{2^{n-k-1}}, 0 \leq k \leq n-1. \quad (6)$$

To obtain the g^nConv output, the result q_n from the last recursive calculation in Equation (3) is fed into the linear projection layer ϕ_{out} . The recursion proceeds for n iterations, during which k is incremented by 1 for each iteration, and pk is updated to p_{k+1} as specified in Equation (4). This enables the n -order information interaction between a feature of the middle layer and its surrounding adjacent region features to be realized. As shown, Figure 4b recognizes the third-order information interaction between the feature and surrounding regions. To improve the calculation efficiency, this paper can compute

the combined adjacent feature $q_{k=0}^{n-1}$ using a deep convolution module f on the GPU for simplified implementation, instead of computing it n times using Equation (4). According to the experiment of paper [9], $g^n\text{Conv}$ shows very competitive performance for COCO object detection [28], ImageNet-1K image classification [29] and ADE20K semantic segmentation [30], which proves the effectiveness of this module. The experimental results in this paper demonstrate that incorporating the $g^n\text{Conv}$ module into the backbone network of YOLOv7 significantly enhances the model's expression ability without adding additional parameters.

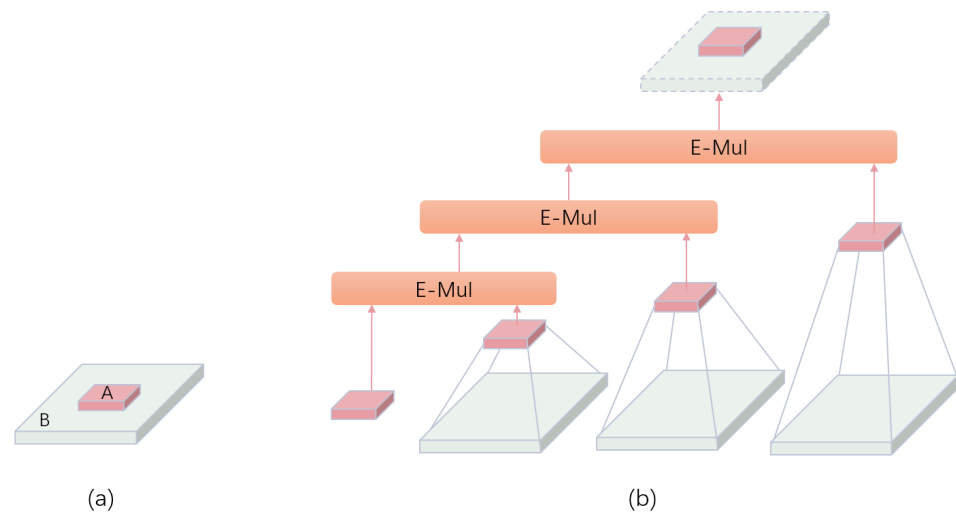


Figure 4. $g^n\text{Conv}$ module realizes spatial information interaction of any stage. (a) Object key feature (area A) and adjacent feature (area B). (b) Three-order information interaction is realized through element multiplication and recursive operation between the target research area and adjacent features.

3.3. Tiny Object Detection Module

Considering that the TinyPerson dataset features ultra-low pixels of foreground objects, which appear at a long distance and a large background, training the model can be difficult and may impact the final detection results. To address this issue, this paper added an additional up-sampling module in the neck layer to improve image resolution and output the result to the fourth detector, which can detect tiny objects on the beach or sea. Furthermore, this study added one detector head and combined it with the other three detector heads to achieve multi-scale detection, while reducing the negative impact of object size changes and making the model more stable. This allows both large and tiny image models to adapt and reduce the negative impact caused by severe object size changes, such as the accurate recognition and positioning of objects. As shown in Figure 3, this paper added an additional up-sampling structure at the end of the sampling structure on the feature pyramid of the neck layer to generate a more expressive feature map. Similarly, to maintain scale matching, this paper added a downsampling structure in the PANet structure [31] to transfer back the more robust positioning features at the lower level to maintain scale matching. This enhances the effect of the multi-scale fusion of features and improves the robustness of the detection scale without significantly increasing the amount of computation. This paper demonstrates that incorporating these modifications into the YOLOv7 backbone network significantly improves the detectability of the model for tiny objects.

3.4. Coordinate Attention

It is well established that the expressiveness of a model is positively correlated with the number of parameters it possesses. Deeper networks tend to be more expressive [32] but also require storing more information during the computation process. YOLOv7 achieves excellent performance and generates a large amount of information for calculation, which may lead to information overload. Therefore, it is necessary to let the network focus on

the area of interest, namely the object area. Attention modules have been widely used in deep learning for this purpose. By focusing on the information that is critical to the task at hand, reducing attention to other information, and even filtering out irrelevant information, the problem of information overload can be addressed, leading to improved efficiency and accuracy of task processing. The coordinate attention module is a plug-and-play module with little computational overhead (see Figure 5 for the network structure diagram). Compared to the transformer method that converts the feature tensor into a single feature vector through two-dimensional global pooling in channel attention [33], the coordinate attention module splits the channel attention into two parallel one-dimensional feature coding processes, effectively integrating spatial coordinate information into the generated attention map.

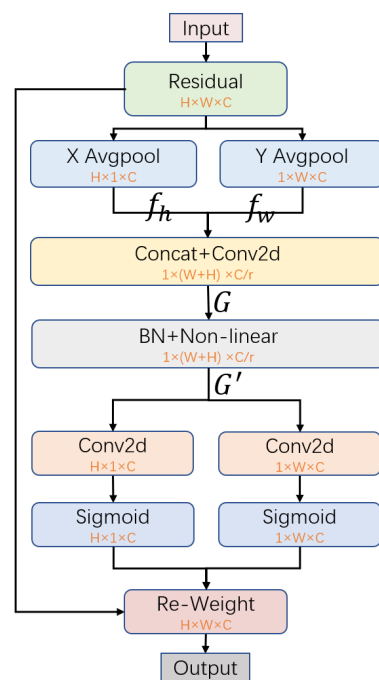


Figure 5. Network structure of coordinate attention mechanism.

As shown in Figure 5, where C represents the number of channels, and r represents the downsampling ratio used to control the size of the module, the CA module can be seen as a computing unit that enhances the expressive power of the network. It takes the tensor with intermediate feature $P = [p_1, p_2, \dots, p_C] \in \mathbb{R}^{HW \times C}$ as input to generate the enhanced representation $Q = [q_1, q_2, \dots, q_C] \in \mathbb{R}^{HW \times C}$ in the same dimension as P . After the residual block [34,35] processing, respectively subject the inputs X and Y to one-dimensional global pooling operation (X represents horizontal direction, Y represents vertical direction), and encode the position information of each channel during the pooling process, then obtain the output f_h at the c -th channel with the height h as shown in Formula (7) and output f_w at the c -th channel with width w as shown in Formula (8). The calculation formula is as follows:

$$f_h = \frac{1}{W} \sum_{0 \leq i \leq W} z_c(h, i), \quad (7)$$

$$f_w = \frac{1}{H} \sum_{0 \leq j \leq H} z_c(j, w). \quad (8)$$

where z refers to the encoding operation that saves the position information in the generated attention map, then splices f_h and f_w proceed with the convolution of 1×1 to obtain the intermediate feature map G in the horizontal and vertical directions. Then, it carries out the normalization and nonlinear transformation operation on G to encode the spatial

information in the vertical and horizontal directions to obtain G' , finally cuts G' back to the two independent directions of h and w , proceeds with the convolution operation of 1×1 to obtain the same number of channels as the input feature map, and finally, the normalization operation is weighted to obtain the final attention weight [36]. This attention operation can distinguish the spatial direction (i.e., coordinates) and generate the coordinate sensing feature map. In short, it not only captures the cross-channel information but also captures the direction sensing and position-sensitive information, which helps the model more accurately locate and identify the objects of interest. By integrating the CA module into the improved YOLOv7, this paper focuses the model's attention on the object area, which improves the model's performance. Specifically, the YOLOv7 MP-2 module is improved to the MP-CA module. For more details, please refer to Figures 3 and 6.

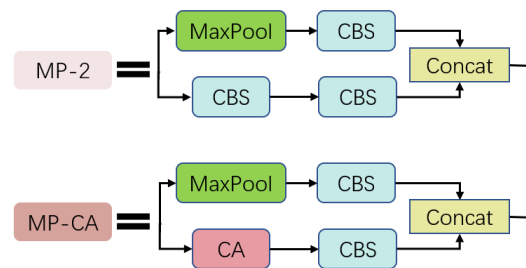


Figure 6. CA focuses on key information by participating in the calculation in MP module.

4. Experiments

4.1. Experimental Setting

The TinyPerson dataset is designed for detecting tiny people and comes from collecting high-resolution videos from different websites, and then collecting images every 50 frames in the video and deleting some repetitive images [2]. The dataset is divided into two parts, including 1610 tagged images and 759 unmarked images, with a total of 72,651 tags. During the training process, adjustments to the model's hyperparameters are typically made using the model's performance on the verification set as a feedback signal. To optimize the performance of the model and prevent overfitting caused by information leakage, this study increased the proportion of the verification set for the TinyPerson dataset. The results of this paper show that a partition closer to 1:1 yields significantly better performance than the traditional 8:2 or 9:1 partition. This study focuses on 1610 tagged images, of which 794 are used as training sets, and the other 816 are used as verification sets and also test sets. To differentiate between people in the water and on land, this paper categorizes the human objects as "sea person" and "earth person". During training, this paper sets the initial learning rate to 0.01, and due to the relatively small size of the user-defined datasets, this paper uses the adaptive moment estimation algorithm (Adam) as the optimization function, which automatically adjusts the learning rate. As training tiny objects requires more time, after testing and comprehensive consideration of training time and GPU memory, this paper sets the number of epochs and batch size to 1000 and 32, respectively. It can be seen from Figure 7 that the model converges after 1000 rounds of training. This paper conducts ablation experiments to improve YOLOv7, using the pre-trained model provided on the official website since many modules of the improved network architecture are the same as the layers in the original YOLOv7 module. This approach saves time and cost and does not significantly affect the experiment's accuracy. The environment used in this experiment is PyTorch 1.10.0, Intel (R) Xeon (R) Platinum 8255C CPU@2.50GHz, and the training and reasoning of all models are conducted on NVIDIA RTX 3090 GPU.

4.2. Evaluation Index

In this paper, the rating indicators choose include AP (average accuracy rate), P (accuracy rate), R (recall rate), and Params (parameter quantity). Set the following rules: TP = "Positive samples are correctly identified as positive samples", TN = "Negative sam-

ples are correctly identified as negative samples”, FP = “Negative samples are incorrectly identified as positive samples”, and FN = “Positive samples are incorrectly identified as negative samples”. We have the following formula:

$$P = \frac{TP}{TP + FP}$$

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

In Formula (9), P represents the proportion of positive correctly predicted samples to all predicted samples, and R represents the proportion of positive correctly predicted samples to all actual samples. IOU is the intersection ratio, which represents the ratio of the intersection area between the detection anchor and the real anchor to their combined area. Its numerical value indicates the accuracy of the object detector’s positioning ability. AP25, AP50, and AP75 represent the average accuracy of IOU thresholds at 0.25, 0.5, and 0.75, respectively. Generally, AP@50:5:95 is used as the primary metric [37] to evaluate the model’s performance. For ease of reference, it is expressed as AP in this paper, which means that the IOU threshold is taken from 0.5 to 0.95 in steps of 0.05, and then the AP mean value under these IOUs is calculated.

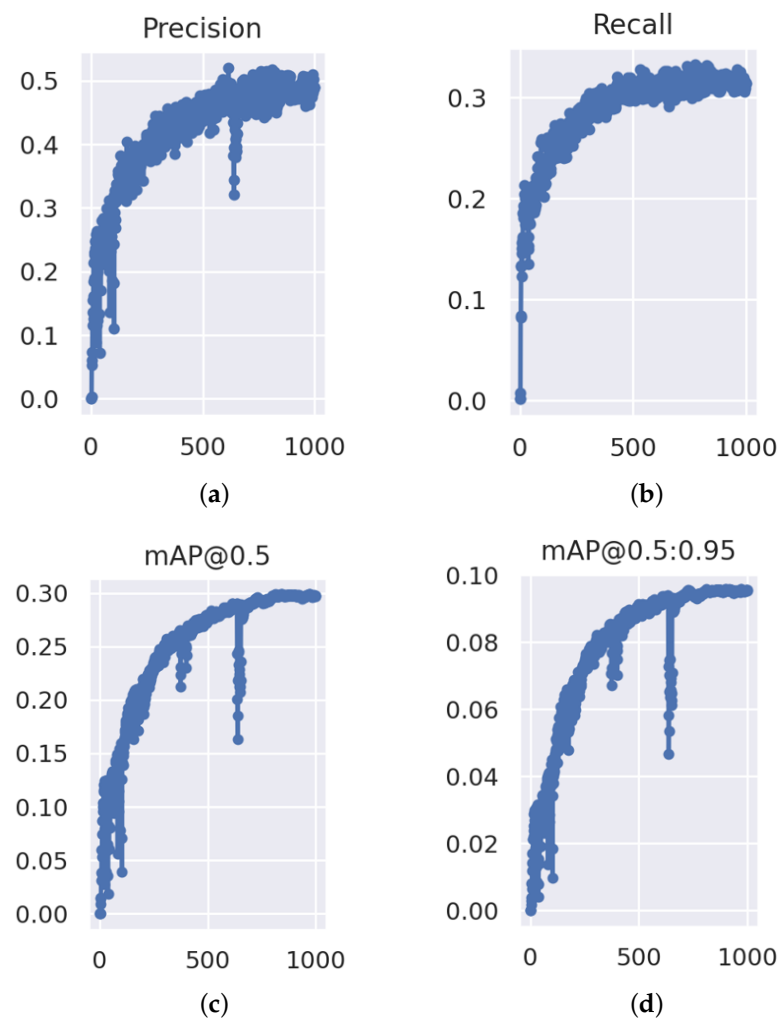


Figure 7. Training curve of TOD-YOLOv7 model. (a) is the curve of precision. (b) is the curve of recall. (c) is the curve of AP@50. (d) is the curve of AP@50:5:95.

4.3. Ablation

This paper analyzed the significance of each component change in the process of improving YOLOv7. Subsequently, this paper conducted local testing of the model after training after each change to verify the impact of adding the g^n Conv module, tiny object detection head, and coordinate attention module on detection performance. All ablation experiments in this paper were carried out on the TinyPerson dataset. The details are shown in Table 1. Each module is added to the baseline network YOLOv7 using gradual improvement. To measure the impact on improving performance, this paper quoted three indicators, parameter quantity, AP50, and AP, for comparison. It can be seen from Table 1 that the g^n Conv module reduced the number of parameters, and the tiny object detection head and coordinate attention module increase the number of parameters, all of which improved the performance.

Effect of g^n Conv module. The baseline network was modified in the backbone, and the number of parameters was reduced by replacing the two-part efficient architecture network ELAN. This did not bring about a performance reduction, which showed the effectiveness of the g^n Conv module. The addition of this module increased AP50 by 1.9%, while reducing the number of parameters by 1.1 M, and increasing AP by 1.1%, proving its effectiveness.

Effect of coordinate attention module. The coordinate attention module enables the network to focus on critical areas. Although it adds some parameter quantity, it does not result in additional calculations. The impact on speed is ignored, and the effect of parameter quantity is balanced. After adding the coordinate attention module, the AP value of the network increased by 0.6%, improving the overall model. This shows that the module enables the network to more accurately locate and identify the objects of tiny people of interest.

Effect of extra prediction head. The addition of the tiny object detection module increased the number of parameters by 2.6M in the model, as the length of the neck layer was increased by 50%. The addition of an upsampling module allowed the model to calculate the image at a higher resolution. This was more effective for the data of most foreground objects, such as TinyPerson, that were less than 20 pixels. It is evident from Table 1 that the increase reached a maximum of 2.6% in the case of AP50.

Table 1. Ablation experiments on the TinyPerson datasets.

Methods	Parms (M)	AP50 (%)	AP (%)
Baseline	34.8	24.9	7.1
+ g^n Conv	33.7(−1.1)	26.8 (+1.9)	8.2 (+1.1)
+ CA	35.3(+1.6)	27.4 (+0.6)	8.8 (+0.6)
+ Head	38(+2.7)	30 (+2.6)	9.5 (+0.7)

4.4. Contrast Experiment

This paper uses the principle of control variables to conduct comparative experiments. Specifically, this paper trains the same dataset using different models to ensure consistency of training parameters and initial hardware training environment. The effectiveness of this work is verified by comparing the final models on the TinyPerson datasets. For the detection task at long distances, this paper selected TPH-YOLOv5, a variant of YOLOv5, to participate in the comparative experiment. TPH-YOLOv5 is designed to operate under the flight conditions of an unmanned aerial vehicle (UAV), where images are captured at a small scale and from far away from the target [25]; this results in blurred and low-resolution images [38]. Furthermore, the UAV captures images with high-density objects that may occlude one another, similar to the scene used in this experiment because the objects it detects already contain pedestrians. Therefore, selecting TPH-YOLOv5 as a model for comparison enables a more accurate evaluation of the performance of the models in this paper. The detailed experimental results are presented in Table 2. As shown in this table, this paper improved the YOLOv7 model, which outperforms other classic one-stage and

two-stage models in terms of AP value under multiple IOU threshold conditions, while having a lower number of parameters. This suggests that the improved network model is more suitable for the scenario of tiny object detection.

Table 2. The comparison of the performance with mainstream object detectors on TinyPerson dataset.

Methods	Image Size	Parms (M)	P (%)	R (%)	FPS	AP25 (%)	AP50 (%)	AP75 (%)	AP (%)
SSD	640 ²	34	24.3	27.6	24	2.4	3.7	1.5	1.8
Faster R-CNN	640 ²	40	15	13.2	15	7.4	15.1	3.2	5.8
YOLOv5	640 ²	6.7	46.2	21.2	86	20.3	20.7	18.2	7.4
TPH-YOLOv5 [25]	640 ²	43.3	48.4	27.4	31	25.1	25.5	20.7	8.2
YOLOv6	640 ²	17.2	45.1	28.3	32	15.2	18.9	3.7	6.8
YOLOv7	640 ²	34.8	47.7	27.5	256	23.6	24.9	22.3	7.1
TOD-YOLOv7 (Ours)	640 ²	38	50.1	32.2	208	28.7	30	27	9.5

5. Conclusions

The TinyPerson dataset used in this paper presents a significant challenge for object detection networks due to its ultra-low pixel objects with less than 20 pixels, long distance, large background, and dense population. To address these challenges, this paper improved upon the most advanced object detector, YOLOv7, by replacing some ELAN modules in the backbone network with recursive gated convolution, adding a tiny object detection module to improve the detection accuracy of small objects, incorporating a coordinate attention mechanism to focus the improved model on critical areas, and utilizing data augmentation during training to enhance performance. Based on these improvements, this paper proposes the TOD-YOLOv7 model with more robust performance in the field of tiny-person detection, which demonstrates significantly superior performance compared to existing mainstream object detectors, achieving an AP of 9.5% in the TinyPerson task. Compared with the original network YOLOv7, the algorithm of this paper exhibits high detection accuracy and robustness, making it feasible for rapidly detecting tiny people in remote and large background scenes through object detection. We hope that this experiment will assist researchers in gaining better insights into the study of tiny people.

Author Contributions: Conceptualization, F.T. and F.Y.; methodology, F.T. and F.Y.; software, F.T.; writing—original draft preparation, F.T.; writing—review and editing, F.T., F.Y. and X.T.; visualization, F.T., F.Y. and X.T.; project administration, F.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research work in this paper was supported by the Science and Technology Project of Hebei Education Department (ZD2019131) and “one province, one university” fund of Hebei University (No. 521000981155).

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset address used in this paper <https://github.com/ucas-vg/PointTinyBenchmark> (accessed on 12 November 2022).

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: A benchmark. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 304–311.
2. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale Match for Tiny Person Detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1246–1254.

3. Jiang, N.; Yu, X.; Peng, X.; Gong, Y.; Han, Z. SM+: Refined Scale Match for Tiny Person Detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 6–11 June 2021; pp. 1815–1819.
4. Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion Loss: Detecting Pedestrians in a Crowd. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7774–7783.
5. Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; Sun, J. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 6448–6457.
6. Zhang, S.; Benenson, R.; Schiele, B. CityPersons: A Diverse Dataset for Pedestrian Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 4457–4465.
7. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
8. Wang, C.; Bochkovskiy, A.; Liao, H.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv: 2207.02696.
9. Rao, Y.; Zhao, W.; Tang, Y.; Zhou, J.; Lim, S.; Lu, J. HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. *arXiv* **2022**, arXiv: 2207.14284.
10. Wang, J.; Chen, Y.; Gao, M.; Dong, Z. Improved YOLOv5 network for real-time multi-scale traffic sign detection. *arXiv* **2021**, arXiv: 2112.08782.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; 2017; pp. 5998–6008.
12. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761. [[CrossRef](#)] [[PubMed](#)]
13. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, 20–26 June 2005; pp. 886–893.
14. Enzweiler, M.; Gavrilu, D.M. Monocular Pedestrian Detection: Survey and Experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2179–2195. [[CrossRef](#)] [[PubMed](#)]
15. Harris, E.; Marcu, A.; Painter, M.; Niranjana, M.; Prügell-Bennett, A.; Hare, J. Fmix: Enhancing mixed sample data augmentation. *arXiv* **2020**, arXiv: 2002.12047.
16. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
17. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
18. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
19. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv: 1804.02767.
21. Bochkovskiy, A.; Wang, C.; Liao, H.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv: 2004.10934.
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2016, Volume 9905, pp. 21–37.
23. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
24. Zhang, X.; Zeng, H.; Guo, S.; Zhang, L. Efficient Long-Range Attention Network for Image Super-Resolution. In Proceedings of the Computer Vision—ECCV 2022 - 17th European Conference, Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part XVII; Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2022, Volume 13677, pp. 649–667.
25. 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 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
26. Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 11966–11976.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021.

28. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision—ECCV 2014—13th European Conference, Zurich, Switzerland, 6–12 September 2014, Proceedings, Part V; Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8693, pp. 740–755.
29. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
30. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene Parsing through ADE20K Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5122–5130.
31. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
32. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV 2018—15th European Conference, Proceedings, Part VII, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11211, pp. 3–19.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
34. Xie, S.; Girshick, R.B.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; pp. 13713–13722.
37. Zhao, H.; Zhang, H.; Zhao, Y. YOLOv7-sea: Object Detection of Maritime UAV Images based on Improved YOLOv7. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023—Workshops, Waikoloa, HI, USA, 3–7 January 2023; pp. 233–238.
38. Liu, T.; Fu, H.Y.; Wen, Q.; Zhang, D.K.; Li, L.F. Extended faster R-CNN for long distance human detection: Finding pedestrians in UAV images. In Proceedings of the IEEE International Conference on Consumer Electronics, ICCE 2018, Las Vegas, NV, USA, 12–14 January 2018; pp. 1–2.

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.