Article

DC-YOLOv8: Small size Object detection algorithm based on camera sensor

Haitong Lou[1], Xuehu Duan[1], Junmei Guo[1], Haiying Liu[1]*, Jason Gu[2], Lingyun Bi[1], Haonan Chen [1]

[1] The School of Information and Automation Engineering, Qilu University of Technology(Shandong Academy of Sciences), Shandong, China.
[2] The School of Electrical and Computer Engineering, Dalhousie University, Halifax, Canada.
* haiyingliu2019@qlu.edu.cn

Abstract: Traditional camera sensors rely on human eyes for observation. However, the human eye is prone to fatigue when observing targets of different sizes for a long time in complex scenes, and human cognition is limited, which often leads to judgment errors and greatly reduces the efficiency. Target recognition technology is an important technology to judge the target category in camera sensor. In order to solve this problem, a small size target detection algorithm for special scenarios was proposed by this paper. Its advantage is that this algorithm not only has higher precision for small size target detection, but also can ensure that the detection accuracy of each size is not lower than the existing algorithm. In this paper, a new down-sampling method was proposed, which could better preserve the context feature information. The feature fusion network was improved to effectively combine shallow information and deep information. A new network structure was proposed to effectively improve the detection accuracy of the model. In terms of accuracy, it is better than: YOLOX, YOLOXR, YOLOv3, scaled YOLOv5, YOLOv7-Tiny and YOLOv8.Three authoritative public data sets were used in this experiment: a) On Visdron data sets (small size targets), DC-YOLOv8 is 2.5% more accurate than YOLOv8. b) On Tinyperson data sets (minimal size targets), DC-YOLOv8 is 1% more accurate than YOLOv8. c) On PASCAL VOC2007 data sets (Normal size target), DC-YOLOv8 is 0.5% more accurate than YOLOv8.

Keywords: YOLOv8, small size targets, target detection, feature fusion.

1. Introduction

As one of the most widely used devices, cameras have been an essential device in various industries and families, such as robotics, monitoring, transportation, medicine, autonomous driving and so on[1–5]. Camera sensor is one of the core sensors of the above requirements, it is composed of lens, lens module, filter, CMOS/CCD, ISP and data transmission part. It works by first collecting images using optical imaging principles and finally performing image signal processing. The application of cameras in traffic, medicine, automatic driving, etc., is crucial to accurately identify the target, so the target recognition algorithm is one of the most important parts in the camera sensor.

Traditional video cameras captured the scene and presented it on the screen, then the shape and type of the object were observed and judged by the human eye. However, human cognitive ability is limited, and it is difficult to judge the category of the object when the camera resolution is too low. When in a complex scene, It will strain the human eye, resulting in the inability to detect some small details. A viable alternative to this problem is to use camera sensors to find areas and categories of interest.

At present, the technology of target recognition through the camera is one of the most challenging topics, and its accuracy and real-time performance are the most important indicators applied in the camera sensor. In recent years, with the ultimate goal of accuracy or real-time, MobileNet[6–8], ShuffleNet[9][10], etc. that can be used on CPU, and ResNet[11], DarkNet[12], etc. that can be used on GPU have been proposed by researchers.

 doi:10.20944/preprints202304.0124.v1

At this stage, the most classical target detection algorithms are divided into Two kinds: two-stage object detection algorithm and One-stage object detection algorithm. The representatives of Two-stage object detection algorithms are R-CNN[13], Fast R-CNN[14], Faster R-CNN[15], Mask R-CNN[16], etc. The representative of one-stage object detection algorithms are YOLO series algorithms[12][17–21], SSD algorithms[22], and so on. The camera sensor pays attention to the real-time performance while ensuring the improvement of accuracy. In complex scenes, multiple targets need to be processed in real time. We focus on the optimization module to enhance the feature extraction ability while lightweight, which ensures the accuracy. So we used a one-stage object detection algorithm. The YOLO series of algorithms is one of the fastest growing and best algorithms so far, especially the novel YOLOv8 algorithm released in 2023 has reached its highest accuracy so far. However, YOLO only solved the target of full size. When the project becomes a special scene with a special size, its performance is not as good as some current small-size object detection algorithms[25][26]. In order to solve this problem, this paper proposed the DC-YOLOv8 algorithm. The detection accuracy of this algorithm has a stable small improvement under normal scale targets and greatly improved the detection accuracy of small targets in complex scenes. The pixels of small targets are small, which will make the detector extract features accurately and comprehensively in the process of feature extraction. Especially in complex scenes such as object overlap, it is more difficult to extract information, so the accuracy of various algorithms for small targets is generally low. In order to greatly improve the detection accuracy of small objects in complex scenes while the detection accuracy of normal-scale objects has a stable and small improvement, The main contributions of the proposed algorithm are as follows:

a) MDC module was proposed to perform downsampling operation(The method of concatenating depthwise separable convolutions, maxpool, and convolutions of dimension size 3×3 with stride=2 is presented), It can supplement the information lost by each module in the downsampling process, making the context information saved in the feature extraction process more complete.

b) The C2f module in front of the detector in YOLOv8 was replaced by the DC module proposed in this paper. A new network structure is formed by stacking DC modules and fusing each small module continuously. It increased the depth of the whole structure, achieves higher resolution without significant computational cost, and was able to capture more contextual information.

c) The feature fusion method of YOLOv8 was improved, which could perfect combine shallow information and deep information, made the information retained in the process of network feature extraction more comprehensive, and the problem of missed detection due to inaccurate positioning was solved.

This paper is divided into the following parts: The second part introduced the reasons for choosing YOLOv8 as the baseline and the main idea of YOLOv8; The third part mainly introduced the improved method of this paper; The fourth part focused on the experimental results and comparative experiments; The fifth part was the conclusion and the direction of subsequent work and improvement.

2. Related Works

2.1. The reason for choosing YOLOv8 as the baseline

YOLO is currently the most popular real-time object detector, which can be widely accepted for the following reasons: a)Lightweight network architecture. b) Effective feature fusion methods. c) The detection results are more accurate. yolov8 is designed to combine the advantages of many real-time object detectors. It still adopts the idea of CSP in YOLOv5[27], feature fusion method (PAN-FPN)[28][29] and SPPF module. Its main improvement is: a) It provided a brand new SOTA model, including P5 640 and P6 1280 resolution object detection networks and YOLACT's instance segmentation model[23]. In order to met the needs of different projects, it also designed models of different scales based on the scaling coefficient like YOLOv5. b) On the premise of retaining the original idea of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1

YOLOv5, the C2f module is designed by referring to the ELAN structure in YOLOv7[21]. c) The detection head part also used the current popular method (separating the classification and detection heads)[30]. Most of the other parts were still based on the original idea of YOLOv5. d) YOLOv8 classification Loss used BCE Loss, The regression Loss was of the form CIOU Loss + DFL, VFL proposes an asymmetric weighting operation[24]. DFL: The position of the box is modeled as a general distribution. Let the network quickly focus on the distribution of the location close to the target location, and make the probability density near the location as large as possible, as shown in formula (1). s_i is the output of sigmod for the network, y_i and y_{i+1} are interval orders, y is label. Compared to the previous YOLO algorithm, YOLOv8 is very extensible. It is a framework that can support previous versions of YOLO, and can switch between different versions, so it is easy to compare the performance of different versions.

$$DFL_{(s_i,s_{i+1})} = -((y_{i+1} - y)\log(s_i) + (y - y_i)\log(s_{i+1})) \tag{1}$$

YOLOv8 used Anchor-Free instead of Anchor-Base. V8 used dynamic TaskAlignedAssigner for matching strategy. It calculates the alignment degree of Anchor-level for each instance using Equation(2), s is the classification score, u is the IOU value, α and β are the weight hyperparameters. It selects m anchors with the maximum value (t) in each instance as positive samples, and selects the other anchors as negative samples, and then trains through the loss function. After the above improvements, YOLOv8 is 1% more accurate than YOLOv5, making it the most accurate detector so far.

$$\mathrm{t} = s^{\alpha} \times \mathrm{u}^{\beta} \tag{2}$$

The key feature of YOLOv8 is that it is extensible. Yolov8 is designed to work with all versions of YOLO and switch between them, making it easy to compare their performance, which is a great benefit for researchers working on YOLO projects. Therefore, YOLOv8 version was selected as the baseline.

2.2. The network structure of YOLOv8

The Backbone part of YOLOv8 is basically the same as that of YOLOv5, and the C3 module is replaced by the C2f module based on the CSP idea. The C2f module learned from the ELAN idea in YOLOv7, and combined C3 and ELAN to form the C2f module, so that YOLOv8 could obtain more abundant gradient flow information while ensured lightweight. At the end of backbone, the most popular SPPF module was still used, and three MaxPools of size 5×5 were passed in serial, and then each layer was concatenation, so as to guarantee the accuracy of targets in various scales while ensuring lightweight simultaneously.

In the Neck part, the feature fusion method used by YOLOv8 is still PAN-FPN, which strengthens the fusion and utilization of feature layer information at different scales. The authors of YOLOv8 used two upsampling and multiple C2f modules together with the final decoupled head structure to compose the Neck module. The idea of decoupling the head in YOLOx, was used by YOLOv8 for the last part of the neck. It combined confidence and regression boxes to achieve a new level of accuracy.

YOLOv8 can support all versions of YOLO, and can switch between different versions at will. It can also run on various hardware platforms (CPU-GPU), which has strong flexibility. Figure 1 shows the YOLOv8 network architecture diagram.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1

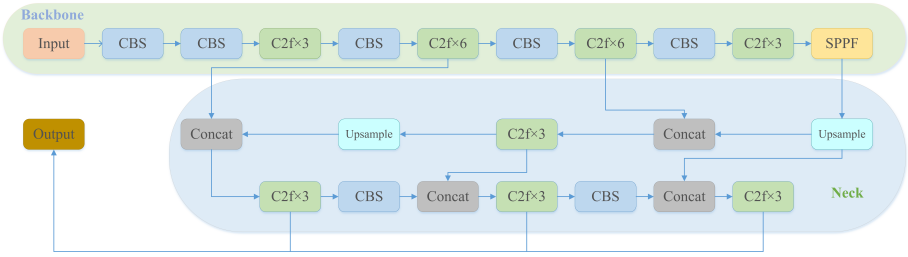


Figure 1. YOLOv8 network structure diagram

3. The proposed DC-YOLOv8 algorithm

YOLOv8 has been very perfect in all aspects, but there are still some problems in the identification of small targets in complex scenes. The reasons for the inaccurate detection of small targets are analyzed as follows: a) When the neural network performs feature extraction, the small-size target is misled by the large-size target, and the features extracted at the deep level lack a lot of small-size target information, which leads to the neglect of small targets in the whole learning process, so the detection effect is poor. b) Compared with normal size, small size objects are easier to overlap with other objects, and are easy to be partially blocked by other size objects, making it difficult to distinguish and locate in the image.

In order to solve the above problems, we proposed a detection algorithm that could greatly improve the detection effect of small size targets on the basis of ensuring the detection effect of normal size. First we proposed the MDC module for downsampling operation, which adopted depth separable convolution, maxpool and convolution of size 3×3 with stride=2 for concatenation. This can fully supplement the information lost in the downsampling process of each item, and can more completely preserve the context information of the feature extraction. Secondly, the feature fusion method was improved, which could better combine shallow information and deep information, so that the information retained in the process of network feature extraction was more comprehensive, and the problem of not detecting the target due to inaccurate positioning and being misled by large-size targets was solved. Finally, a DC module (depthwise separable convolution + convolution of size 3×3) that was constantly stacked and fused was proposed to form a new network structure, and it was replaced by the C2f module in front of the detection head. This method increases the depth of the whole structure and obtains higher resolution without increasing significant computational cost. It can capture more context information and effectively improve the problem of low detection accuracy caused by object overlap.

3.1. A modified efficient downsampling method

The downsampling method used in this paper mainly contains three parts, which were Maxpool, Depthwise separable convolution and convolution module of size 3×3.

Common downsampling methods are generally a separate 3×3 convolutional module or Maxpool. Maxpool can alleviate the over-sensitivity of the convolutional layer to the location, which can better improve the robustness of the target. However, Maxpool will filter out the information that it thinks is not important when performing downsampling, and the information of small-size targets is easy to be misled or even masked by the information of large-size targets. Therefore, when Maxpool is carried out, part of the target information will be automatically filtered out, and only the important information that is considered by itself will be left, which reduces the resolution and has a great impact on the final accuracy. In order to achieve lightweight, Google proposed MobileNet, and proposed Depthwise separable convolution (DW) for the first time. DW convolution has smaller volume and less calculation, which makes it one third of the calculation of regular convolution in the training process. However, it will also lose a lot of useful information while reducing the amount of calculation. The convolution module of size 3×3 has a high degree of nonlinearity and can represent more complex functions. Small convolution kernels can extract small features, so every researcher is willing to use ordinary convolution

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1

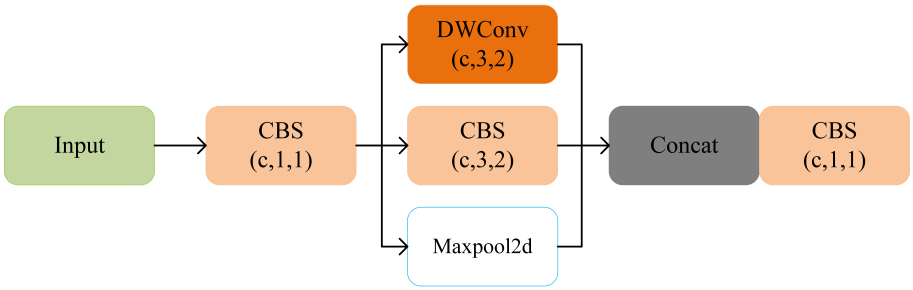
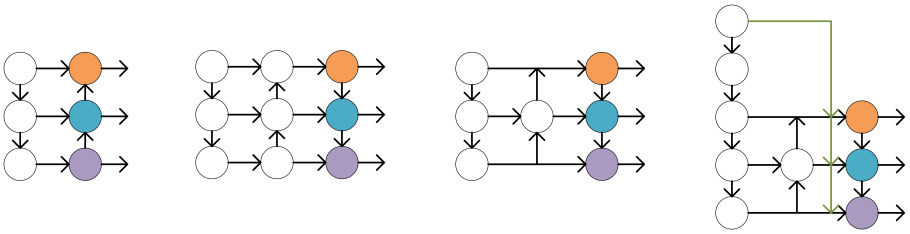


Figure 2. Downsampling method



(a) FPN (b) PAN (c) Yolov8 (d) DCS-Yolov8

Figure 3. Comparison of DCS feature fusion method and other feature fusion methods

for downsampling operation. However, in the whole process of network extraction, there are many down-sampling operations used, so the amount of calculation is not ignored.

Therefore, this paper first used the convolution of size 1×1 for dimensionality reduction, and then used the convolution of size 3×3 for downsampling, which would reduce a lot of calculation. During this operation, the Maxpool layer and Depthwise separable convolution were concatenated. This can fully supplement the information lost in the down-sampling process of each item, and can more completely preserve the context information during feature extraction. After many experiments, it is proved that the MDC module is more efficient than the downsampling method of YOLOv8 original, and the specific structure is shown in the Figure 2.

3.2. Improved feature fusion method

When feature extraction is carried out in special scenarios, small-size targets are easy to be misled by normal-size targets, resulting in less and less information until it disappears completely. Moreover, the problem of poor target positioning in the whole network has always existed. Therefore, we improve the feature fusion method.

Observing the entire network structure diagram of YOLOv8, it can be seen that although the feature fusion method has retained both shallow information and deep information, the target positioning information is mainly existed in the shallower position. In the process of feature extraction, the information that is not obvious is automatically deleted by the network, so a lot of important information for small size objects is deleted in the shallowest layer, which is also the main reason for the poor detection of small objects. The improved feature fusion method in this paper was focused on solving this problem. In the first layer of feature extraction, the size of the picture was changed to 80×80, 40×40, 20×20 by Maxpool. It was then concatenated with the outputs of different scales separately. The reason of using Maxpool for downsampling is that Maxpool can extract the main location information and filter out other useless information in the process of downsampling, and has a very low amount of calculation. Through the feature extraction of each layer, what is missing is the most original feature information, so only using Maxpool for downsampling operation can meet the needs of this paper. The specific structure is shown in the Figure 3

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1

3.3. The proposed network structure

In order to solve the problem of losing a lot of important information due to being misled by large-size objects in the feature extraction process, a deeper network architecture (DC) was proposed in this paper. The DC module adopts the idea of DenseNet and VOVNet[31][32]. It gathers each important module in the last layer at the same time, and gathers the important information of the previous layers in each layer, which can well avoid the problem of information loss, and ensure that the normal size information and small size information can be well preserved. However, in the process of the experiment, we found that the results are not as good as we imagined. After many experiments, the previous feature extraction was found to use a single convolution module, and the convolution module cannot extract complete information well. The parallel concatenation of convolution and Depthwise separable convolution can learn from each other and improve the learning ability and stability of the network.

Via the experiment, the most stable state could be achieved by replacing the C2f module in front of the head detector with the DC module. The specific structure is shown in the Figure 4

After the improvement of the above three improved methods, the new network learning ability has been greatly improved and enhanced. The DC-YOLOv8 network structure is shown in Figure 5

4. Experiments

The new algorithm was trained and tested on Visdrone dataset to improve each stage, and compared with YOLOv8. In order to verify that this algorithm could improve the detection accuracy of small size targets without reducing the accuracy of other scale targets, comparative experiments was carried out on PASCAL VOC2007 dataset and Tinperson dataset. Finally, we selected complex scene pictures in different scenarios to compare the detection effects of the proposed algorithm and YOLOv8 algorithm in actual scenes.

After many experiments, it can be known that the algorithm basically iterates 120 and then begins to converge. According to the hardware facilities and multiple experimental attempts, we set the following parameters: batch size=8, epoch=200.

4.1. Experimental platform

The system used for the experiments in this paper is Windows11, and the system hardware facilities: 16G RAM, NVIDIA GTX3070 GPU, Intel i512400f CPU. Software platform: torch 1.12.1+cu113, Anaconda.

4.2. Valuation index

Evaluation metrics: Mean average precision (mAP), average precision (AP), precision (P), and recall (R). The formulas for P and R are as follows:

$$P = \frac{TP}{(TP + FP)} \tag{3}$$

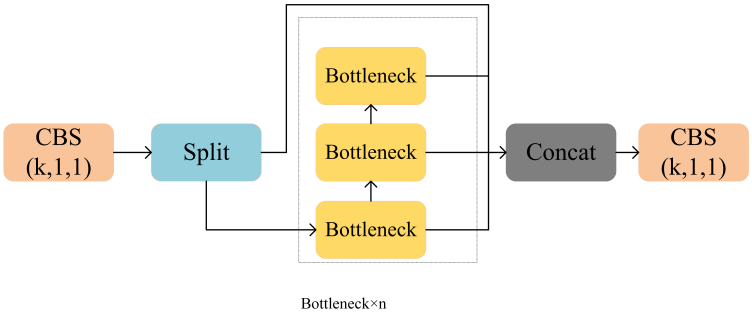
$$R = \frac{TP}{(TP + FN)} \tag{4}$$

TP is the number of correctly predicted bounding boxes, FP is the number of incorrectly judged positive samples, and FN is the number of undetected targets.

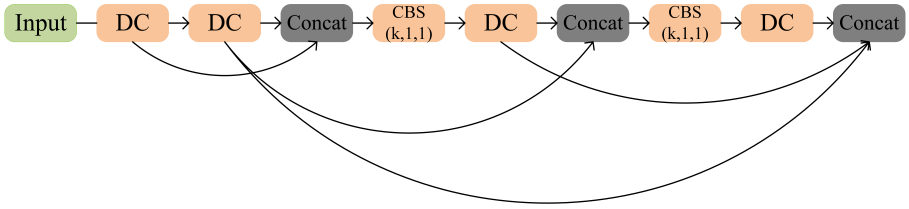
Average Precision (AP) is the average accuracy of the model. mean Average Precision (mAP) is the average value of the AP. K is the number of categories. The formulas for AP and mAP are as follows:

$$AP = \int_0^1 p(r)dr \tag{5}$$

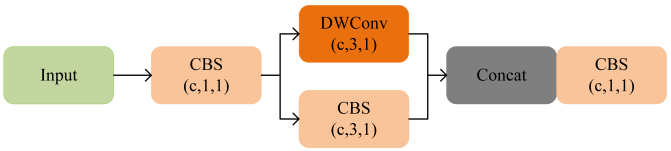
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1



（a）C2f



（b）DC Network Structure



（c）DC

Figure 4. Figure (a) shows the C2f module, which is designed by referring to the idea of C3 module and ELAN, so that YOLOv8 can obtain more abundant gradient flow information while ensuring lightweight. Figure (b) shows the network structure proposed in this paper. It not only adopts the idea of DenseNet and VOVNet, but replaces the original convolution with a parallel cascade of convolutions and depthwise separable convolutions. Figure (c) is the basic block in the network architecture, which is composed of convolutions and depthwise separable convolutions

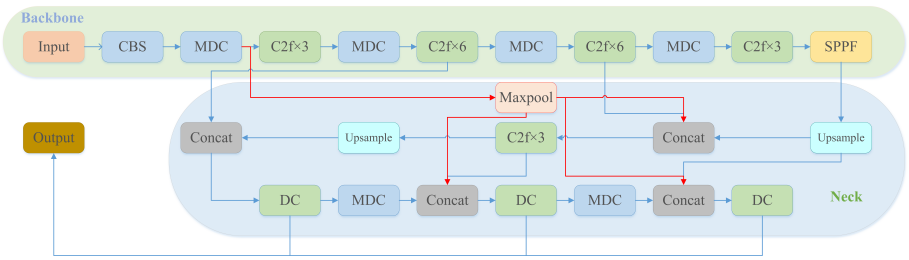


Figure 5. Network structure diagram of DC-YOLOv8

$$mAP = \frac{1}{k}\sum_{i=1}^{k} AP_i \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1

4.3. Experimental results analysis

In order to verify the detection effect of the improved method in this paper on small-size targets in each stage, we conducted ablation experiments on each stage on the Visdrone dataset and compared it with YOLO v8s. The Visdrone dataset was collected by the AISKYEYE team at Tianjin University, China. The dataset is acquired by UAV and has a wide coverage, it is collected in different scenes, weather, light, so there are numerous small size targets in complex environments. This dataset also provides some attributes such as scene visibility, object class, and occlusion. The Visdron data set is extensive and authoritative, and this data set conforms to the content studied in this experiment in all aspects. So we used this data set for control experiments.

In order to clearly show the authenticity of the experiment, this experiment uses mAP0.5, mAP0.5:0.9 as the evaluation index. The test results are shown in TABLE I.

Table 1. Algorithm comparison at each stage

Detection algorithm	Module			Result			
	MDC	Feature fusion	DC	mAP0.5	mAP0.5:0.95	P	R
YOLOv8				39	23.2	50.8	38
DC-YOLOv8	✓			39.5	23.5	51.2	38.8
DC-YOLOv8	✓	✓		40.3	24.1	51.8	39.4
DC-YOLOv8	✓	✓	✓	41.5	24.7	52.7	40.1

TABLE I showed that for the detection of small size targets in complex scenes, the improved algorithm has a certain improvement in each stage. And the recall rate is improved by 2%, which means that there is a lot of room for improvement. It can be proved that the three methods improved in this experiment are obviously effective: a) The improvement of the down-sampling method can fully supplement the information lost in the down-sampling process, and can save the context information during feature extraction more completely. b) The improvement of feature fusion method effectively prevented the problem of small targets being ignored in the whole learning process due to location information. c) The improvement of the network structure effectively solved the problem of losing a lot of important information due to being misled by large-size objects in the feature extraction process. The experimental results showed that the improvement of the algorithm in each stage can improve the learning ability of the model.

In order to compare the detection effect of different types of objects in the DC-YOLOv8 algorithm, we recorded the mAP of 10 kinds of objects in the Visdrone dataset, and the specific results were shown in Figure 6. From the results, we can see that there are four categories whose recognition accuracy is higher than the average level of the whole dataset. The modified algorithm has a steady improvement on larger objects such as car, and a large improvement on smaller objects such as tricycle,bicycle,awning-tricycle, etc.

4.4. Comparison of experiments with different sizes objects

Second set of experiments was a comparison experiment of different sizes, and the data sets used were Pascal VOC2007 and Tinyperson data sets. Pascal VOC2007 is one of the most authoritative datasets and its object types are divided into four categories: Vehicle, House-hold, Animal, and Person, and 20 small types. The resolution of Tinyperson data set is very low, basically less than 20 pixels, and it is for small target detection of long distance and complex background. There are two categories of TinyPerson, sea person and earth person. Each algorithm (YOLOv3, YOLOv5, YOLOv7, YOLOv8, DC-YOLOv8) is tested on three data sets at the same time and recorded for comparison, as shown in the TABLE II. The total number of iterations was set by us to 200 rounds, and recorded their mAP0.5 and mAP0.5:0.95, respectively. From the TABLE II, we can conclude that the experimental

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1

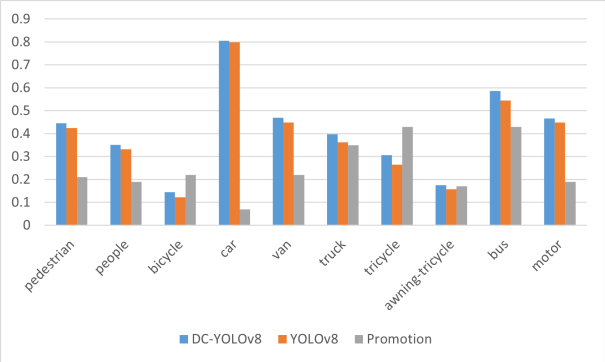


Figure 6. Comparing the 10 categories of YOLOv8 and DC-YOLOv8, blue is the result of DC-YOLOv8 proposed in this paper, orange is the result of YOLOv8, and gray is the accuracy of the difference between the two algorithms

results of DC-YOLOv8 are significantly higher than those of other classical algorithms in the experiments of small-size targets and even for extremely small-size targets, and DC-YOLOv8 is also slightly higher than other algorithms in the experiments of normal-size targets. In order to facilitate subsequent verification, the weight file was saved with the highest mAP value during the experiment.

Table 2. Detection effect of data sets with different scales

Detection algorithm	Dataset			Result	
	Visdrone	VOC	Tinyperson	mAP0.5	mAP0.5:0.95
YOLOv3	✓			38.8	21.6
YOLOv5	✓			38.1	21.7
YOLOv7-tiny	✓			30.7	20.4
YOLOv8	✓			39	23.2
DC-YOLOv8	✓			41.5	24.7
YOLOv3		✓		79.5	53.1
YOLOv5		✓		78	51.6
YOLOv7-tiny		✓		69.1	42.4
YOLOv8		✓		83.1	63
DC-YOLOv8		✓		83.5	64.3
YOLOv3			✓	18.5	5.79
YOLOv5			✓	18.3	5.81
YOLOv7-tiny			✓	16.9	5.00
YOLOv8			✓	18.1	6.59
DC-YOLOv8			✓	19.1	7.02

The reasons why DC-YOLOv8 algorithm was better than other algorithms were analyzed: a) Most of the feature fusion methods used by classical algorithms are FPN+PAN, and small-size targets are easy to be misled by normal-size targets in the process of feature extraction layer by layer, resulting in the loss of most of the information. The feature fusion method of DC-YOLOv8 could fuse the shallow information in the final information result well, and effectively avoided the problem of information loss in shallow layers. b)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1

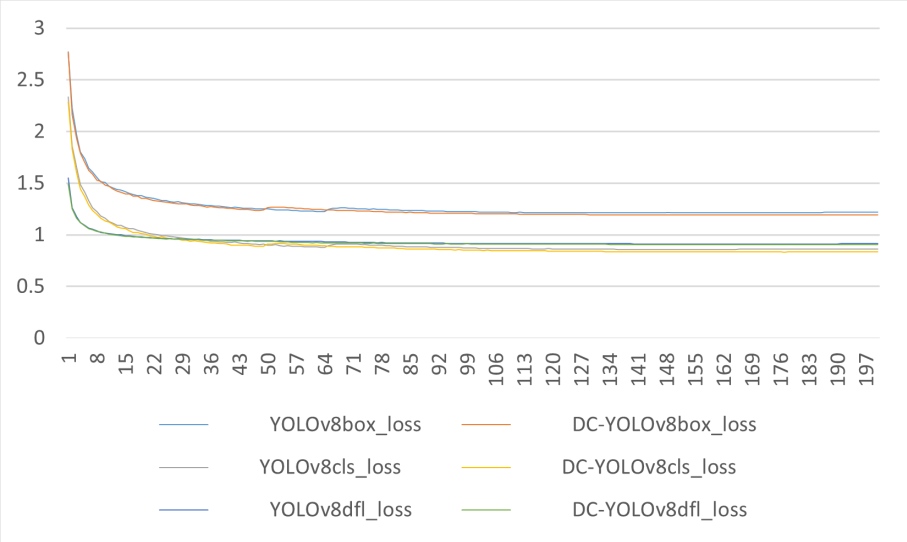


Figure 7. Comparison of class loss between DC-YOLOv8 and YOLOv8

(a) Test results of YOLOv8 with inference time of 12ms (b) Test results of DC-YOLOv8 with inference time of 12ms

Figure 8. Experimental comparison of complex scenes in life

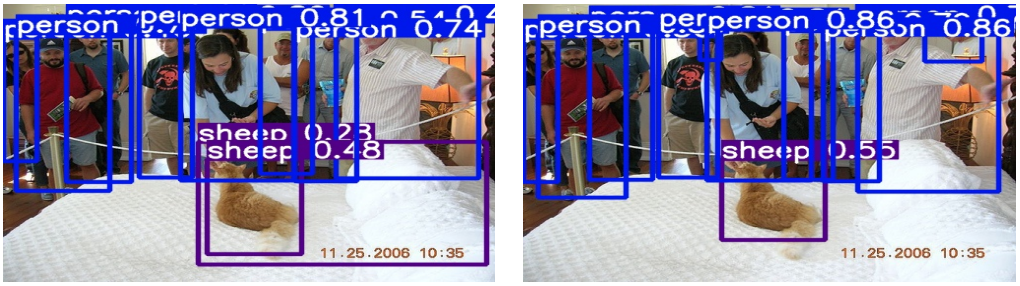
Unimportant information will be automatically ignored in feature extraction, and small-size target pixels will be ignored when extracting features, resulting in reduced accuracy. However, DC-YOLOv8 adopts the idea of DenseNet and VOVNet, which uses a deeper network and is able to learn more detailed information. The class loss of DC-YOLOv8 and YOLOv8 is shown in Figure 7. We could see from the Figure 7 that DC-YOLOv8 had a certain degree of reduction in the three category losses of box loss, class loss, and dfl loss.

In order to intuitively see the detection effect of DC-YOLOv8, two sets of complex scene graphs were selected for testing. The weight files with the highest accuracy of DC-YOLOv8 and YOLOv8 was retained by us and used for test comparison. Image selection criteria: complex scenes, with targets under various sizes, with overlapping targets. The difference between DC-YOLOv8 and YOLOv8 can be clearly seen through the above requirements.

Among them, Figure 8 showed the comparison in complex life scene (the highest weight file of Visdrone dataset is used). Figure 9 showed the detection comparison of normal-sized objects (the highest weight file of Pascal VOC2007 dataset is used). It can be seen from Figure 8 to Figure 9 that DC-Yolov8 is higher than Yolov8 in terms of both the number of detected targets and the accuracy of detected targets.

In the first group of comparison experiments, the images of complex living scenes in Visdrone dataset was selected by us, which had complex scenes, more interference and overlap. Figure 8 showed that we could see that YOLOv8 has false detection in the leftmost part of the picture due to the dark color. There are false detections at the middle tent, mainly due to overlapping objects. The right bike is misdetected due to the complex and overlapping environment nearby. On the far right, it is not detected because of incomplete

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0124.v1



(a) Test results of YOLOv8 with inference time of 12ms (b) Test results of DC-YOLOv8 with inference time of 12ms

Figure 9. Comparison diagram of the normal size target experiment

information. In the middle position, there is a false detection because of overlap. It can be seen that although YOLOv8 has many advantages, there are still some problems under small-size targets. In contrast, DC-YOLOv8 can accurately detect the right target when only partial information is available, and accurately detect the target in complex and overlapping scenes without false detection or missed detection. It can be seen from Figure 8 that the detection effect of DC-YOLOv8 is better than YOLOv8 when the target size is small.

For the second set of comparison experiments, the images with multiple people overlapping in the PASCALVOC2007 dataset were selected. Figure 10 showed that the two people overlap in the position of the middle door, and we can only see the head of the person behind because of the occlusion of the person in front. In this case, the YOLOv8 detector will fail to detect the person leaking only the head. In the position of the cat, there is a false detection (detecting the human arm as a cat) because the color of the human arm is similar to that of the cat. In the case of severe overlap on the rightmost side, YOLOv8 does not detect the person behind. In contrast, in the case of overlap, DC-YOLOv8 accurately detected the person near the door and the person to the far right, and there was no false detection due to similar colors. It can be seen from Figure. 9 that DC-YOLOv8 also outperforms YOLOv8 in the detection of normal-sized objects.

5. Conclusions

This paper proposes a small size object detection algorithm based on camera sensor, different from traditional camera sensor, we combine camera sensor and artificial intelligence. Then, some problems in the newly released YOLOv8 and the existing small-size object detection algorithms are analyzed and solved. New feature fusion methods and network architectures are proposed. It greatly improves the learning ability of the network. The test and comparison are carried out on Visdrone dataset, Tinyperson dataset and PASCAL VOC2007 dataset. Through analysis and experiments, the feasibility of each part of the optimization is proved. DC-YOLOv8 outperforms other detectors in both accuracy and speed. Small targets in various complex scenes are easier to capture.

In the future, we will continue to conduct in-depth research on camera sensors, and strive to achieve the goal of being able to outperform existing detectors in detection accuracy in various sizes as soon as possible.

Author Contributions: Junmei Guo, Haiying Liu and Jason Gu have given me technical guidance and writing method guidance as my instructors, Xuehu Duan, Lingyun Bi and Haonan Chen have done experiments and writing together as my classmates

Acknowledgments: This work was supported by QLUTGJHZ2018019

Sample Availability: The source code for the experiments is available at the author.

References

1. M. Y. Zou, J. J. Yu, Y. Lv, B.Lu, W. Z. Chi and L. n. Sun. A Novel Day-to-Night Obstacle Detection Method for Excavators based on Image Enhancement and. IEEE Sensors Journal, 2023, pp. 1–11.

2. H. Liu and L. L. Member. Anomaly detection of high-frequency sensing data in transportation infrastructure monitoring system based on fine-tuned model. IEEE Sensors Journal, 2023, pp. 1–9.
3. F. Zhu, Y. Lv, Y. Chen, X. Wang, G. Xiong, and F. Y. Wang. Parallel Transportation Systems: Toward IoT-Enabled Smart Urban Traffic Control and Management. 2020, IEEE Transactions on Intelligent Transportation Systems., vol. 21, no. 10, pp. 4063–4071.
4. J. Thevenot, M. B. Lopez, and A. Hadid. A Survey on Computer Vision for Assistive Medical Diagnosis from Faces.2018, IEEE journal of biomedical and health informatics, vol. 22, no. 5, pp. 1497–1511.
5. A. D. Abadi, Y. Gu, I. Goncharenko, and S. Kamijo. Detection of Cyclist's Crossing Intention based on Posture Estimation for Autonomous Driving. 2023, IEEE Sensors Journal, pp. 1–1.
6. A. G. Howard, M. l. Zhu, B. Chen, D. Kalenichenko, W. j. Wang, T. Weyand, M Andreetto, H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications.
7. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520.
8. A. Howard, W. Wang, G. Chu, L. Chen, B. Chen, and M. Tan. Searching for MobileNetV3 Accuracy vs MADDs vs model size. 2019, Proceedings of the IEEE/CVF international conference on computer vision, pp. 1314–1324.
9. X. Zhang, X. Zhou, M. Lin, and J. Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. 2018, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6848–6856.
10. N. Ma, X. Zhang, H. T. Zheng, and J. Sun. Shufflenet V2: Practical guidelines for efficient cnn architecture design. 2018, Proceedings of the European conference on computer vision, vol. 11218 LNCS, pp. 122–138.
11. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016, Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 2016-Decem, pp. 770–778,
12. J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. 2018, arXiv preprint arXiv:1804.02767.
13. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587.
14. R. Girshick. Fast R-CNN.2015, Proceedings of the IEEE international conference on computer vision, vol. 2015 Inter, pp. 1440–1448.
15. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. 2015, 28, Advances in neural information processing systems.
16. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. 2020, Proceedings of the IEEE international conference on computer vision, vol. 42, no. 2, pp. 386–397.
17. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. 2016, Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 2016-Decem, pp. 779–788.
18. J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. 2017, Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR 2017, vol. 2017-Janua, pp. 6517–6525.
19. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. 2020, [Online]. Available: http://arxiv.org/abs/2004.10934.
20. C. Li et al., YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. 2022, [Online]. Available: http://arxiv.org/abs/2209.02976.
21. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2022, pp. 1–15.
22. W. Liu et al., SSD: Single shot multibox detector. 2016, Computer Vision–ECCV, vol. 9905 LNCS, pp. 21–37.
23. D. Bolya, C. Zhou, F. Xiao, and Y. Lee Jae. Yolact: Real-time Instance Segmentation. 2019, Proceedings of the IEEE/CVF international conference on computer vision, pp. 9157–9166.
24. Y. Cao, K. Chen, C. C. Loy, and D. Lin. Prime Sample Attention in Object Detection. 2020, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11583-11591.
25. H. Liu, X. Duan, H. Chen, H. Lou, and L. Deng. DBF-YOLO:UAV Small Targets Detection Based on Shallow Feature Fusion.2023, IEEJ Transactions on Electrical and Electronic Engineering, doi: 10.1002/tee.23758.
26. H. Liu, F. Sun, J. Gu, and L. Deng. SF-YOLOv5: A Lightweight Small Object Detection Algorithm Based on Improved Feature Fusion Mode. 2022, Sensors, vol. 22, no. 15, pp. 1–14.
27. C. Y. Wang, H. Y. Mark Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh. CSPNet: A new backbone that can enhance learning capability of CNN. 2020, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, vol. 2020-June, pp. 1571–1580.
28. T. Lin, R. Girshick, K. He, B. Hariharan, S. Belongie. Feature Pyramid Networks for Object Detection. 2017, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125.
29. S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path Aggregation Network for Instance Segmentation. 2018, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8759–8768.
30. Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. YOLOX: Exceeding YOLO Series in 2021. 2021, pp. 1–7.
31. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. 2017, Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR 2017, vol. 2017-January, pp. 2261–2269.

 doi:10.20944/preprints202304.0124.v1

32. Y. Lee, J. W. Hwang, S. Lee, Y. Bae, and J. Park. An energy and GPU-computation efficient backbone network for real-time object detection. 2019, Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 2019-June, pp. 752–760.