applied sciences

MDPI

Article

Lightweight You Only Look Once v8: An Upgraded You Only Look Once v8 Algorithm for Small Object Identification in Unmanned Aerial Vehicle Images

Zhongmin Huangfu † and Shuqing Li *

School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China; hfzmylh@163.com

* Correspondence: lishuqing@stu.ncwu.edu.cn

† Current address: School of Software, North China University of Water Resources and Electric Power, Zhengzhou 450046, China.

Abstract: In order to solve the problems of high leakage rate, high false detection rate, low detection success rate and large model volume of small targets in the traditional target detection algorithm for Unmanned Aerial Vehicle (UAV) aerial images, a lightweight You Only Look Once (YOLO) v8 algorithm model Lightweight (LW)-YOLO v8 is proposed. By increasing the channel attention mechanism Squeeze-and-Excitation (SE) module, this method can adaptively improves the model's ability to extract features from small targets; at the same time, the lightweight convolution technology is introduced into the Conv module, where the ordinary convolution is replaced by the GSConv module, which can effectively reduce the model computational volume; on the basis of the GSConv module, a single aggregation module VoV-GSCSPC is designed to optimize the model structure in order to achieve a higher computational cost-effectiveness. The experimental results show that the LW-YOLO v8 model's mAP@0.5 metrics on the VisDrone2019 dataset are more favorable than those on the YOLO v8n model, improving by 3.8 percentage points, and the computational amount is reduced to 7.2 GFLOPs. The LW-YOLO v8 model proposed in this work can effectively accomplish the task of detecting small targets in aerial images from UAV at a lower cost.

Keywords: YOLO v8; unmanned aerial vehicle; small targets; target detection; attention mechanism

Citation: Huangfu, Z.; Li, S. Lightweight You Only Look Once v8: An Upgraded You Only Look Once v8 Algorithm for Small Object Identification in Unmanned Aerial Vehicle Images. *Appl. Sci.* **2023**, *13*, 12369. https://doi.org/10.3390/app132212369

Academic Editor: Alessandro Lo Schiavo

Received: 19 October 2023
Revised: 12 November 2023
Accepted: 14 November 2023
Published: 15 November 2023

1. Introduction

With the rapid development of UAV technology, UAVs are widely used in the fields of security monitoring, emergency rescue, agriculture and environmental detection, logistics and warehouse management, construction and construction site monitoring by virtue of their lightweight and fast characteristics. Recently, great progress has been made in target detection. In particular, with the development of large-scale visual datasets and the increase in computational power. Deep neural networks (DNN) [1], especially convolutional neural networks (CNN) [2], have demonstrated record-breaking performance in computer vision tasks including target detection [3–5].

Although deep-learning-based approaches have made great progress in target detection, the problem of off-target detection still exists in UAV. Since these algorithms are mostly used in target detection scenarios under natural perspectives, it is difficult to achieve satisfactory results when detecting small targets such as pedestrians, bicycles, and other small targets in the data of UAV aerial photography images. Compared with target detection in ordinary scenes, UAV aerial images have the following characteristics: (1) dense distribution of targets with a small pixel occupancy; (2) complex UAV aerial images with a large difference in the angle of the same type of targets; and (3) UAV hardware limitations, which make it difficult to provide sufficient arithmetic power and space. These characteristics make target detection task throwing from the UAV viewpoint very challenging.

Aiming at the above problems, this work proposes a lightweight LW-YOLO v8 model suitable for small target detection based on the YOLO [6] v8 model. By adding the channel attention mechanism SE module [7], the model adaptively performs feature extraction on the key regions of small targets to improve the detection accuracy; introducing a lightweight convolution technique in the Conv module [8], replacing the ordinary convolution with the GSConv [9] module, which can effectively reduce the amount of model computation; on the basis of the GSConv module, a single aggregation module VoV-GSCSPC [10] is designed to optimize the model structure to achieve higher computational cost-effectiveness. This method maintains the information exchange between channels with low time loss, which reduces the amount of model computation while improving the detection accuracy.

2. Related Work

For target detection in UAV aerial images, deep learning-based target detection algorithms [11] are subdivided into two categories, namely, one-stage target detection and two-stage target detection. Typical one-stage target detection algorithms are You Only Look Once (YOLO) [12], where the image is processed only once to obtain the bounding box coordinates and class regression probabilities. Typical two-stage target detection algorithms are R-CNN [13], Fast R-CNN [14], Faster R-CNN [15], R-FCN [16], SPP-Net [17], NAS-FPN [18], etc. It first generates a series of candidate boxes called regions of interest through an algorithm, and then a convolutional neural network divides the regions of interest for classification. Li et al. [19] developed a lightweight convolutional neural network structure called WearNet for automatic scratch detection on contact sliding parts such as metal forming. The structure is capable of gradually extracting and learning discriminative features for surface scratch classification, which minimizes the number of network layers and parameters while ensuring high classification accuracy and recognition rate for steel surface scratches, and provides significant advantages in terms of model size and classification depth. However, the performance of the scratch detection model is different in a variety of different scenarios, and the work does not adequately assess the generalization ability of the model. Jawaharlalnehru et al. [20] proposed an improved YOLO algorithm, which improved the detection results from the perspective of model efficiency by using target frame dimensional clustering, pre-training of classification network, training of multi-scale detection, and changing the filtering rules of candidate frames, which verified the feasibility of image localization problem in target detection. However, the research data available in the existing research is small, and the workload of data sample production is difficult to justify. Li et al. [10] introduced a new lightweight convolutional technique GSConv module, and designed a single aggregation module VoV-GSCSP module to replace the ordinary bottleneck module, which improves the inference speed while increasing the detection accuracy, this method proposes a novel lightweight target detector architecture, Slim-neck, which improves the accuracy and robustness of target detection by introducing global and self-attention module into a conventional target detector. However, its introduction of the attention module increases the computational complexity and the number of parameters, which can have an impact on the real-time performance of the target detector. Zhu et al. [21] adding the transformer encoder module to replace traditional CNN prediction heads to better capture target features and contextual information, and introduces a self-attention mechanism and positional encoding, enhancing the model's perception of the spatial distribution and relative position of targets. This method refreshed the detection record of the VisDrone2021 dataset, but introduced additional computational and parameter complexity, which had an impact on the real-time performance and efficiency of the target detector. Zhang et al. [22] improved the problem of the low detection rate of small targets, such as pedestrians, by means of data enhancement to the dataset of aerial images from UAV through dense cropping and the addition of the localized attention module, which effectively improved the detection rate of small targets such as pedestrians from the UAV viewpoint. However, this method requires a large amount of annotated data for training, which poses difficulties in obtaining data for certain application scenarios. Chen et al. [23]

proposed a new target detection method based on a deep separable multi head network structure. This method reduces the number of network parameters while reducing the model volume, and can quickly and accurately detect small targets in drone aerial images. However, for certain specific types of targets, such as small and irregularly shaped targets, the detection accuracy of this method may decrease. Zhang et al. [24] proposed a method based on multi-scale hollow convolutional receptive field, which considers target information at multiple scales and can more comprehensively describe target features, improving the accuracy of small target detection tasks in drone aerial images. This method introduces the idea of dense connections, connecting multiple feature maps at different scales, enabling the network to fully utilize information at different scales and improving the robustness of object detection. However, this method requires a certain amount of computing resources for model training and inference, which may pose certain challenges for unmanned aerial vehicle devices with limited computing resources. Chen et al. [25] effectively improved the network's detection ability for drone aerial images by adding feature fusion modules and residual hole convolution. The modified model has the advantages of high occlusion target recognition rate and low false alarm rate. However, the improved YOLO v5s requires a certain amount of computing resources and hardware support, which poses certain challenges for resource limited drone devices. Yuan et al. [26] proposed an improved YOLO v8 model, which introduced more data augmentation techniques, used multi-scale feature fusion, and improved loss functions to quickly and accurately detect and recognize fish at a lower cost. However, this paper only used commercial fishing vessel electronic monitoring datasets for experimental verification, and did not test them in other scenarios such as datasets from the perspective of drones. Therefore, the transferability of this method needs to be verified.

3. Methods

We proceed with our work according to the sequence of the workflow diagram in Figure 1. In it, Figure 1a represents the data processing stage, Figure 1b represents the model optimization stage, while Figure 1c represents the result analysis stage.

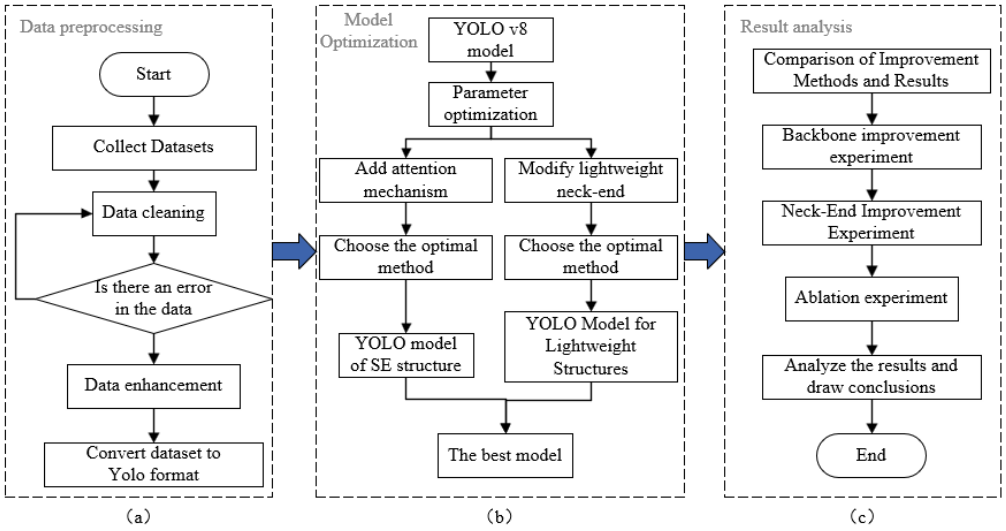


Figure 1. Workflow diagram. (**a**) The data processing stage. (**b**) The model optimization stage. (**c**) The result analysis stage.

3.1. LW-YOLO v8 Network Model

Due to the high detection efficiency of the YOLO series detectors, they are more widely used in industry. In this work, we use the YOLO v8n network model to accomplish small target detection in UAV aerial images. The YOLO v8n network model [27] is the latest version of the YOLO family of algorithms, and its network structure is mainly divided into four parts: the input for data enhancement, the BackBone network for image feature

extraction, the Neck-end for feature fusion, and the decoupled header output that separates the classification from the detection header. The structure of the LW-YOLO v8 model proposed in this work is shown in Figure 2.

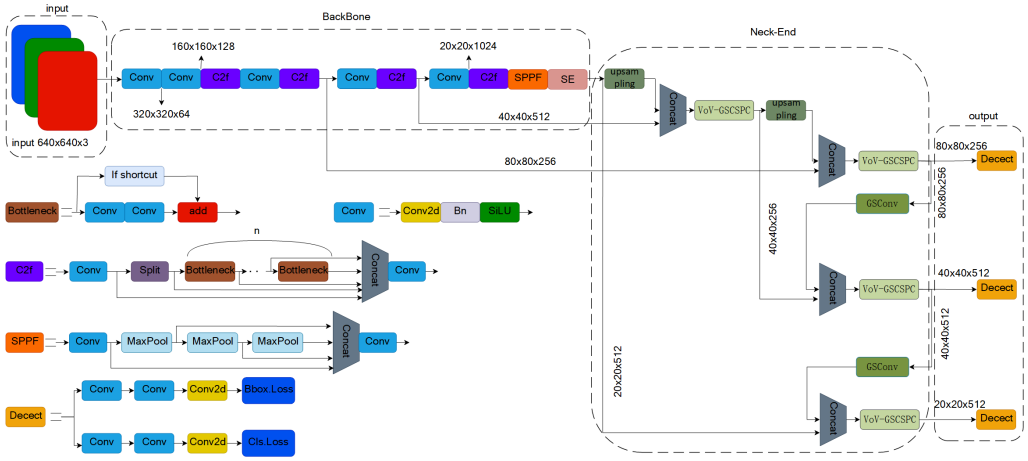


Figure 2. LW-YOLO v8 Network Model.

The YOLO v8 model in BackBone uses the Spatial Pyramid Pooling Fast (SPPF) feature pyramid network, which reduces the computational effort to some extent and increases the receptive field, but the non-multi-scale prediction makes it weak in terms of small target detection. Small targets often have less pixel information in the image, making it difficult to accurately localize and classify them. The channel attention mechanism, SE module, can adaptively adjust the importance of each channel in the feature map by learning the degree of attention of different channels in the image, which makes the model better for feature extraction of key regions of small targets by adding SE module.

The Neck-end YOLO v8n model is modeled after the Efficient Layer Aggregation Networks (ELAN) structure of YOLO v7 [28], and adopts the Cross Stage Partial Network Fusion (C2f) module, which is connected across layers by more branches to form a neural network model with stronger feature representation capability while enriching the model gradient flow; however, the model structure is relatively large and deep, which requires longer time for convergence and training, and the training is slower and more computationally intensive compared to other target detection algorithms. The lightweight convolutional technique GSConv module and the thin-neck detector VoV-GSCSPC module capture local and global structural features by aggregating multilayer neighbor information, and process the input feature maps in groups, and each group independently performs convolutional operations and parallel computation. This computational approach improves the parametric and computational efficiency of the model and reduces the computation by reducing the number of parameters and computational load.

3.2. SE-BackBone Network

Due to the dense distribution of targets and small target scales in the UAV aerial image dataset, when using the baseline model YOLO v8n for target detection and recognition, the model has a weak feature extraction capability for small targets and a small detection accuracy, which cannot satisfy the detection accuracy requirements of UAV aerial images. To solve the above problems, the fixed-length feature vectors Q_1 generated from the multi-scale input images processed by the feature pyramid pooling SPPF module are used as inputs to the SE module of the channel attention mechanism, which focuses on the information that is more critical to the current task among the many inputs and suppresses the useless information through adaptive learning of the relevance and importance of the feature channels, so as to improve the characterization ability of the network model. In this work, the channel attention mechanism SE module is applied to UAV aerial image feature extraction; as shown in Figure 3, the SE module is embedded into the BackBone

network to form a new feature extraction BackBone network, SE-BackBone, and the SE module enhances the differential information extraction of UAV aerial images by weighting the importance of the feature extracted images.

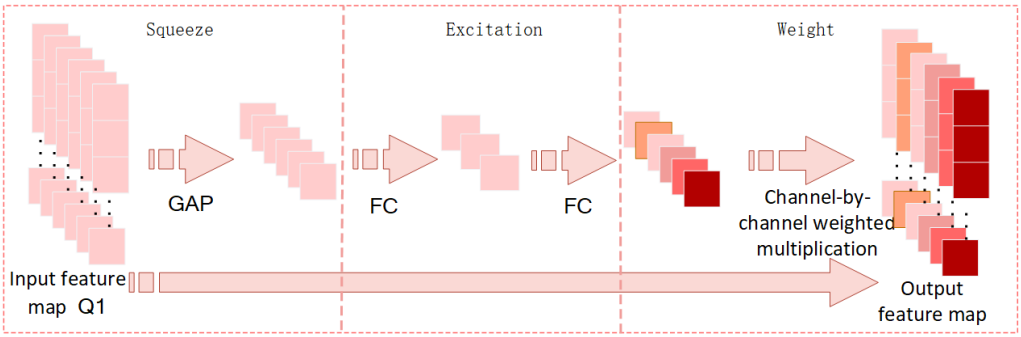


Figure 3. SE Module.

When the SE module is introduced into the network model, the network model training can be targeted to enhance the learning of target feature regions, which helps to improve the model's detection accuracy of small targets in UAV aerial images. In the network model, the output feature map Q_1 of the SPPF layer is passed into the SE module as the input features, which are sequentially subjected to the squeezing operation of the global average pooling (GAP) and the excitation operation composed of the two full connectivity layers (FC), and the feature maps Q_2 outputted from the squeezing process and the channel weighted feature maps obtained from the excitation process are compared with the outputs of the squeezing process by the weighting operation one by one. The weight operation multiplies the output feature map Q_3 from the extrusion process with the channel weighted feature map obtained from the excitation process channel by channel to re-calibrate the importance of the input features. The SE module mapping relationship is shown in Equations (1)–(3).

$$L_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{1}$$

$$L_{ex}(s, W_0) = \theta(g(s, W_0)) = \theta(W_2 \varphi(W_1 s)) \tag{2}$$

$$L_{we} = (u_c, e_c) = e_c u_c \tag{3}$$

where L_{sq}, L_{ex} and L_{we} represent the squeeze operation, the excitation operation and the weight operation, respectively; H and W are the height and width of the input feature map Q_1, respectively; the value of i is in the range of [0, H], and the value of j is in the range of [0, W]; u_c represents each feature channel; s represents the true number after the Squeeze operation; φ represents the Relu nonlinear activation function; W_0 represents the total parameters of the FC operation; W_1, W_2 and g represent the parameters of the FC operation, respectively; θ represents the Sigmoid activation function performs the weight normalization operation on each feature map; and e_c represents the weights obtained from the Excitation operation.

3.3. Lightweight Neck-End

Generalized approaches to enhance the learning ability of CNN, such as DensNet [29], VoVNet [30] and CSPNet [31], were investigated, and lightweight Neck-end were designed based on the theoretical foundations of these approaches. When using the baseline model YOLO v8n for target detection and recognition in the target detection process of UAV aerial images, it cannot meet the higher cost of computational resources and faster detection speed. Instead of ordinary convolution (Figure 4a), the lightweight convolution technique GSConv module is used in this work to preserve information exchange between channels with low time loss. The time complexity of convolution calculation is usually defined

by GFLOPs, and the time complexity (unbiased) of GSConv is: $\text{Time}_{GSConv} \sim O[W \cdot H \cdot K_1 \cdot K_2 \cdot \frac{C_2}{2}(C_1 + 1)]$, where W is the width of the output feature map; H is the height of the output feature map; $K_1 \cdot K_2$ is the size of the convolutional kernel; C_1 is the number of channels for each convolutional kernel, which is also the number of channels for the input feature map; and C_2 is the number of channels for the output feature map. The advantages of the GSConv module are evident in lightweight detectors, which capture local and global structural features using aggregated multilayer neighborhood information, and use a uniform mixing (Shuffle) operation to infiltrate information generated by standard convolution into information generated by depth separable convolution, and is specifically neck designed for channel maximization and size minimization. But if GSConv module is used at all stages of the model, the network layer of the model will be deeper, which will intensify resistance to data flow and significantly increase inference time. Therefore, we only use GSConv module at the neck of the network model. At this stage, using GSConv module to process the connected feature map is correct: there is less redundant and repetitive information, and compression is not required.

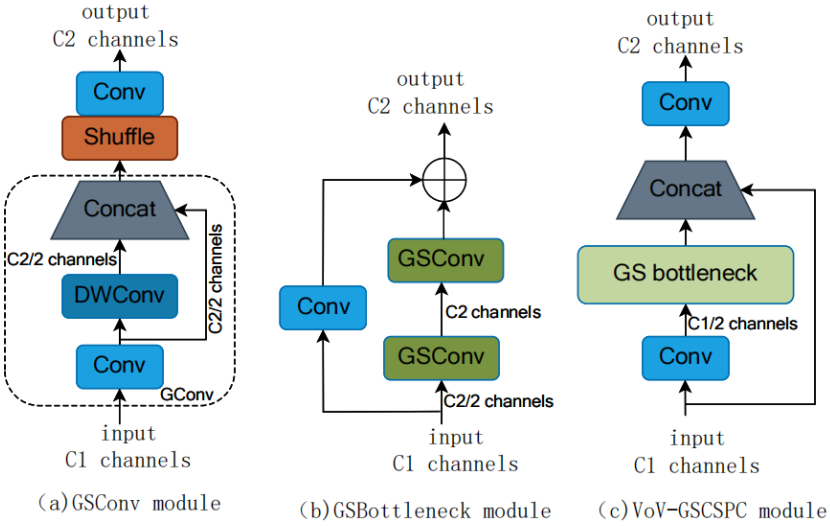


Figure 4. LW-YOLO v8 Network Model. (**a**) The lightweight convolution technique GSConv module. (**b**) The GSBottleneck module. (**c**) The single aggregation module VoV-GSCSPC .

In this work, the GSBottleneck module and the single aggregation module VoV-GSCSPC are further used on top of GSConv module, as shown in Figure 4b,c, where GSConv module is used instead of the normal convolution operation at the Neck-end, and the single aggregation module VoV-GSCSPC is used in place of the original normal bottleneck module C2f. The lightweight module design reduces the model computation while ensuring the accuracy of model detection.

4. Experiments

The experimental environment of this work is as follows: Linux Ubuntu 16.04.4 LTS operating system, NVIDIA GeForce RTX 3090 GPU with 24GB video memory, YOLO v8n model as the baseline model, hyper-parameter batch size is set to 8, the training epochs is set to 150, the optimizer is set to SGD, and the initial learning rate is set to 0.01.

4.1. Evaluation Indicators

The computational volume was used to measure the execution efficiency of the LW-YOLO v8 network model in terms of giga floating-point operations per second (GFLOPs), the smaller the computational volume represents the less computational resources occupied

and the faster the model computation. Mean average precision (mAP) is used to measure the detection accuracy of the model, and the calculation formula is shown in Equation (4).

$$P_{\mathrm{mA}} = \sum P_{\mathrm{A}} / N \tag{4}$$

where P_{mA} is the mAP value; N is the total number of categories; P_A is the area under the curve enclosed by recall as the horizontal axis and precision as the vertical axis. mAP@0.5 is the average precision mean at an Intersection over Union (IOU) threshold of 0.5.

4.2. Visualization Results

From the experimental results, it can be seen that the improved model LW-YOLO v8 in this work outperforms the baseline model YOLO v8n and other mainstream target detection models in the detection effect of small targets. Figure 5 shows the comparison graph of the detection effect between the LW-YOLO v8 model and the baseline model YOLO v8n.

(**a**) Result 1

(**b**) Result 2

(**c**) Result 3

Figure 5. Visual results comparisons between YOLO v8n model (left) and LW-YOLO v8 model (right). (**a**) Comparison chart of leakage before and after improvement. (**b**) Comparison chart of misdetection before and after improvement. (**c**) Comparison chart of small target leakage before and after improvement.

From the comparison graph in Figure 5, it can be seen that the LW-YOLO v8 model in this work detects more objects than the baseline model, especially small target objects such as pedestrians, bicycles, motorbikes, etc., and the detection confidence of the same target object is also improved. From Figure 5a, it can be seen that when the objects are very small, the YOLO v8n model ignores the pedestrian target and has a missed detection situation, while the GSConv module adopted by LW-YOLO v8 model maximizes the retention of hidden connections between each channel, thereby improving the semantic information of the context and avoiding the occurrence of missed detection of small targets such as pedestrians. From Figure 5b, it can be seen that YOLO v8n model has a misdetection

situation, while the SE module used by LW-YOLO v8 model allows the network to learn the importance of each feature channel in a fine-grained manner, thus increasing the network's sensitivity to key feature extraction for small targets with no false detections which misclassifies the red bricks as trucks, whereas LW-YOLO v8 does not have a misdetection. From Figure 5c, it can be seen that that for small targets in the distance, the YOLO v8n model misses many objects, while the LW-YOLO v8 model introduces the global spatial information of the feature map into the local space, so that the feature map can better capture the global information of the target, and detect the small targets in the distance, and have a better detection effect.

4.3. Comparison of Improvement Methods and Results

In order to investigate the effect of the proposed improvement methods in this work, the network structure and parameters of the YOLO v8n model were optimized and adjusted, and the VisDrone2019-DET-train dataset was used as the training dataset, and the VisDrone2019-DET-val dataset was used as the validation dataset. Among them, the VisDrone2019 dataset was collected by the AISKYEYE team in the Machine Learning and Data Mining Laboratory of Tianjin University. The benchmark dataset consists of 288 video clips, consisting of 261,908 frames and 102,09 static images, captured by various drone cameras, covering a wide range of locations (from 14 different cities in China thousands of kilometers apart), environments (urban and rural), objects (pedestrians, vehicles, bicycles, etc.), and densities (sparse and crowded scenes). There were no crossover data in the training and validation dataset datasets, so as to experimentally compare the effect of various improvement methods.

4.3.1. BackBone Improvement Experiment

The BackBone network is improved by adding the channel attention mechanism SE module, the abbreviation is GAM for Global Attention Mechanism module [32], the abbreviation is SelfAttention for Self-attention Mechanism module [33], and the abbreviation is CBAM for the region- and channel-based attention mechanism module [34], respectively, to the 10th layer of the BackBone network (starting from layer 0) of the YOLO v8n model.

Their performance comparisons are shown in Table 1; adding different types of attention mechanisms can improve the detection accuracy of the model to different degrees under the same experimental conditions, the addition of the channel attention mechanism SE module brings the highest detection benefit. The channel attention mechanism SE module is used to improve the detection accuracy by calculating the weight coefficients of each channel and scaling and fusion of the features; its mAP@0.5 increased by 3.2%, but providing accurate feature representations while making the network computationally intensive, and its computation increased by 3.9 GFLOPs.

Table 1. Comparison of model performance after BackBone improvement.

Model	mAP@0.5/%	mAP@0.95/%	Layer	GFLOPs Computation
YOLO v8n	32.5	18.8	168	8.1
YOLO v8n+Self	34.9	21.2	179	10.3
YOLO v8n+CBAM	35.1	20.4	176	10.2
YOLO v8n+GAM	35.3	20.4	179	11.7
YOLO v8n+SE	35.7	20.9	175	12.0

4.3.2. Neck-End Improvement Experiment

The results of the impact of different lightweight structures on the model accuracy and computation are demonstrated by using different methods for improvement at the Neck side of the YOLO v8n model. Different convolution methods at the Neck side and the C2f module constructed by this convolution method.

The performance comparisons are shown in Table 2, and the experimental results show that the DSConv module [35], DCNv2 module [36] and GSConvmodule can reduce

the computational amount of the model to different degrees. Due to the limitations of the local receptive field and spatial attention of deep convolution, DSConv+DSConv2D module mainly focuses on the dynamic sampling of feature points, which cannot capture the detailed features of small targets well, leading to a decrease in detection accuracy. The DCNv2+C2f_DCN module improves the modeling capability of object deformation and spatial relationship by introducing deformable convolution, but the introduction of deformable convolution increases the number of model parameters. The GSConv module used in this work can increase the range of the sensing field by introducing a grid sampling mechanism and adapts to small targets with different sizes and attitudes. The GSConv+VoV-GSCSPC module achieves better results, reduces the computational volume of the model by about 11%, and improves the detection accuracy of the improved model by 1 percentage point.

Table 2. Comparison of model performance after Neck improvement.

Model	mAP@0.5/%	Params (M)	GFLOPs Computation
YOLO v8n	32.5	3.01	8.1
DSConv+SConv2D	32.1	2.39	6.6
DCNv2+C2f_DCN	32.7	3.20	7.4
GSConv+VoV-GSCSPC	33.5	2.82	7.2

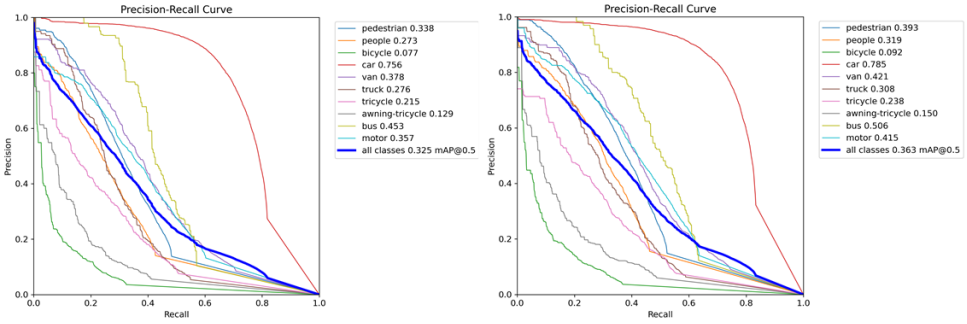
4.4. Ablation Experiment

Using the trained network structure for target detection, the ViSDrone2019-DET-val dataset used as the validation sample data are shown in Table 3: mAP@0.5 is improved by 3.2% by adding the SE module of the channel attention mechanism in the backbone network; the lightweight convolution technique GSConv module is used instead of the ordinary convolution in the Neck side of the feature fusion, and a single-time convergence module VoV-GSCSPC module instead of ordinary C2f module, the mAP@0.5 of LW-YOLO v8 model is improved by 1 percentage point and the computation amount is reduced by 11.11%, which significantly reduces the computational resources; when using the SE module of the channel attention mechanism and the lightweight structure GSConv+VoV-GSCSPC module at the same time, it can be performed to reduce the computation amount and improve the mAP at the same time.

Table 3. Ablation experiment results.

SE	GSConv+VoV-GSCSPC	Precision	Recall	mAP@0.5/%	Params(M)	GFLOPs Computation
✘	✘	43.7	32.5	32.5	3.01	8.1
✔	✘	47.5	35.5	35.7	4.45	12.0
✘	✔	44.6	33.6	33.5	2.82	7.2
✔	✔	47.6	35.5	36.3	2.83	7.2

Target detection using the trained network structure is shown in Figure 6, where Figure 6a shows the experimental results of the baseline model YOLO v8n and Figure 6b shows the experimental results of the LW-YOLO v8 model. It can be seen that the detection accuracy of the LW-YOLO v8 model is better than that of the baseline model YOLO v8n in each category, in which the ten categories of pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor are improved by 0.055, 0.046, 0.015, 0.029, 0.043, 0.032, 0.023, 0.021, 0.053, 0.058. The value of IOU is set to 0.5, and the value of mAP@0.5 for the LW-YOLO v8 model is 0.363, which is higher than that of the baseline model YOLO v8n.



(a) YOLO v8n (b) LW-YOLO v8

Figure 6. Precision-Recall curve. (**a**) PR plot for target detection using YOLO v8n baseline model. (**b**) PR curves for target detection using the improved LW-YOLO v8 model.

Combining the experimental results, using the SE module of the channel attention mechanism and the lightweight structure GSConv+VoV-GSCSPC module, the final LW-YOLO v8n model outperforms the baseline model YOLO v8n in terms of accuracy and computation, with an improvement of 3.8% in mAP@0.5 and a reduction in computation by 0.9% GFLOPs on the VisDrone2019-DET-val validation dataset.

4.5. Comparative Experiments on the Detection of Different Models

In order to validate the performance of the improved YOLO v8 network model, DetNet59 [37], CornerNet [38], Fast R-CNN [39], CenterNet [40], Mixed YOLO v3-LITE [41], and YOLO v8n models are selected for the comparative experiments of target detection, and the results of the experiments are shown in Table 4. In terms of mAP@0.5 metrics, the LW-YOLO v8 model proposed in this work achieves a significant improvement, compared to DetNet59 by 21%, CornerNet by 18.9%, Fast R-CNN by 14.6%, CenterNet by 10.1%, Mixed YOLO v3-LITE by 7.8%, and YOLO v8n model improved by 3.8%. The maximum AP is achieved on the eight categories of pedestrian, people, car, van, truck, tricycle, bus, and motor with 39.3%, 31.9%, 78.5%, 42.1%, 30.8%, 23.8%, 50.6%, and 41.5%, respectively. The experimental results show that the LW-YOLO v8 model proposed in this work can effectively improve the network's detection ability for small targets and performs well in the target detection task of UAV aerial images.

Table 4. Detection results of different algorithms on VisDrone2019 dataset.

Networks	Pedestrian	People	Car	Van	Truck	Tricycle	Bus	Motor	mAP@0.5/%
DetNet59 [21]	15.3	4.1	36.1	17.3	20.9	13.5	26	10.9	15.3
CornerNet [22]	20.4	6.6	40.9	20.2	20.5	14	24.4	12.1	17.4
Fast R-CNN [23]	21.4	15.6	51.7	29.5	19	13.1	31.4	20.7	21.7
CenterNet [24]	22.6	20.6	59.7	24	21.3	20.1	37.9	23.7	26.2
Mixed YOLO v3-LITE [25]	34.5	23.4	70.8	31.3	21.9	15.3	40.9	32.7	28.5
YOLO v8n	33.5	27.3	75.5	37.9	27.7	21.6	45	35.5	32.5
LW-YOLO v8	39.3	31.9	78.5	42.1	30.8	23.8	50.6	41.5	36.3

5. Conclusions

In this work, for the problem of poor small target detection, high false detection rate and high leakage rate in UAV scenarios, we propose a lightweight detection algorithm model LW-YOLO v8 based on the YOLO v8n model. Firstly, we add the channel attention mechanism SE module to this model to adaptively learn the importance of each feature channel to enhance the attention to the small target objects and to fully capture the global information, to adapts the problem of feature inconspicuousness of small targets in UAV scenarios; secondly, the use of lightweight convolution GSConv module instead of ordinary convolution maximally preserves the hidden connections between each channel, thus

enriching the semantic information of the context, and designing a kind of one-shot aggregation module VoV-GSCSPC module to reduce the complexity of the computation and the network structure, and to reduce the amount of model computation while improving the detection accuracy. The experimental results on the UAV dataset VisDrone2019 show that the LW-YOLO v8 model proposed in this work outperforms other mainstream models for small target detection, mAP@0.5 reaching 36.3%, and the model's computational volume is 7.2 GFLOPs, which can effectively complete the task of detecting and recognizing small targets in aerial images from UAV, and in the future it has a positive role to promote in the fields of security monitoring, emergency rescue, agriculture and environmental detection, logistics and warehouse management, construction and site monitoring. Although the LW-YOLO v8 model is better than the baseline model in target detection in UAV aerial images, it still needs to be improved and optimized. For example, the accuracy and speed of target detection are still to be improved, and it has not yet met the requirements of practical applications, and we will carry out more in-depth research in the areas of model volume, detection speed, and number of parameters to contribute to the development of target detection in the field of UAV aerial images in the future. We will perform more in-depth research on model volume, detection speed, number of parameters, etc. In order to contribute to the future development of target detection in UAV aerial images.

Author Contributions: Z.H.: Conceptualization, Methodology, Resources, and Writing—review and editing, formal analysis, Validation and reviewing. S.L.: Investigation, Data collection, Data interpretation, data curation, Software, Validation and Visualization, formal analysis, and Writing—original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Funds of key Research Project of Higher Education Institutions in Henan Province, China (No.21A520026), and the Funds of Science and Technology Research Projects in Henan Province, China (No.202102210378).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The public data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare that they have no conflict of interest to report regarding the present study.

References

1. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
2. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882.
3. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I.
4. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef]
5. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
6. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A review of yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
7. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
8. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv* **2014**, arXiv:1412.6806.
9. Hu, J.; Wang, Z.; Chang, M.; Xie, L.; Xu, W.; Chen, N. Psg-yolov5: A paradigm for traffic sign detection and recognition algorithm based on deep learning. *Symmetry* **2022**, *14*, 2262. [CrossRef]
10. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by gsconv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:2206.02424.

11. Bekkerman, I.; Tabrikian, J. Target detection and localization using mimo radars and sonars. *IEEE Trans. Signal Process.* **2006**, *54*, 3873–3883. [CrossRef]
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
14. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; Volume 28. [CrossRef]
16. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 5–10 December 2016; Volume 29.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
18. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7036–7045.
19. Li, W.; Zhang, L.; Wu, C.; Cui, Z.; Niu, C. A new lightweight deep neural network for surface scratch detection. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 1999–2015. [CrossRef] [PubMed]
20. Jawaharlalnehru, A.; Sambandham, T.; Sekar, V.; Ravikumar, D.; Loganathan, V.; Kannadasan, R.; Khan, A.A.; Wechtaisong, C.; Haq, M.A.; Alhussen, A.; et al. Target object detection from unmanned aerial vehicle (uav) images based on improved yolo algorithm. *Electronics* **2022**, *11*, 2343. [CrossRef]
21. 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, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
22. Zhang, X.; Feng, Y.; Zhang, S.; Wang, N.; Mei, S. Finding nonrigid tiny person with densely cropped and local attention object detector networks in low-altitude aerial images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2022**, *15*, 4371–4385. [CrossRef]
23. Chen, W.; Jia, X.; Zhu, X.; Ran, E.; Hao, X. Target detection in unmanned aerial vehicle images based on DSM-YOLO v5. *J. Comput. Eng. Appl.* **2023**, *59*, 226–233.
24. Zhang, R.; Shao, Z.; Wang, J. Multi-scale void convolutional target detection method for unmanned aerial vehicle images. *J. Wuhan Univ. Inf. Sci. Ed.* **2020**, *45*, 895–903.
25. Xu, C.; Peng, D.; Yu, G. Real time object detection of unmanned aerial vehicle images based on improved yolov5s. *Optoelectronics* **2022**, *49*, 210372-1.
26. Yuan, H.C.; Tao, L. Detection and identification of fish in electronic monitoring data of commercial fishing vessels based on improved Yolov8. *J. Dalian Ocean. Univ.* **2023**, *38*, 533–542.
27. Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-time flying object detection with Yolov8. *arXiv* **2023**, arXiv:2305.09972.
28. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
29. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
30. Lee, Y.; Hwang, J.-W.; Lee, S.; Bae, Y.; Park, J. An energy and gpu-computation efficient backbone network for real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
31. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. Cspnet: A new backbone that can enhance learning capability of cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Washington, DC, USA, 14–19 June 2020; pp. 390–391.
32. Caron, E.; Feuerwerker, L.C.M.; Passos, E.H. Gam, apoio e cuidado em caps ad. *Polis Psique* **2020**, *10*, 98–121. [CrossRef]
33. Lin, Z.; Feng, M.; Santos, C.N.D.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A structured self-attentive sentence embedding. *arXiv* **2017**, arXiv:1703.03130.
34. 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.
35. Nascimento, M.G.D.; Fawcett, R.; Prisacariu, V.A. Dsconv: Efficient convolution operator. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5148–5157.
36. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
37. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Detnet: A backbone network for object detection. *arXiv* **2018**, arXiv:1804.06215.

38. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
39. Yu, W.; Yang, T.; Chen, C. Towards resolving the challenge of long-tail distribution in uav images for object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA , 3–8 January 2021; pp. 3258–3267.
40. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
41. Zhao, H.; Zhou, Y.; Zhang, L.; Peng, Y.; Hu, X.; Peng, H.; Cai, X. Mixed yolov3-lite: A lightweight real-time object detection method. *Sensors* **2020**, *20*, 1861. [CrossRef] [PubMed]