

Article

A Dense Small Object Detection Algorithm Based on a Global Normalization Attention Mechanism

Huixin Wu, Yang Zhu * and Liuyi Wang

School of Information Engineering, North China University of Water Resources and Electric Power, No. 136 Jinshui East Road, Zhengzhou 450046, China; wuhuixin@ncwu.edu.cn (H.W.); 18336720944@163.com (L.W.)

* Correspondence: zhuyang0311@163.com

Abstract: To address the challenges of detecting a large number of objects and a high proportion of small objects in aerial drone imagery, we proposed an aerial dense small object detection algorithm called Global Normalization Attention Mechanism You Only Look Once (GNYL) based on the Global Normalization Attention Mechanism. In the backbone network of GNYL, we embedded a GNAM (Global Normalization Attention Mechanism) that explores channel attention features and spatial attention features from input features in a concatenated manner. It utilizes batch normalization's scale factors to suppress irrelevant channels or pixels. Furthermore, the spatial attention sub-module introduces a three-dimensional arrangement with a multi-layer perceptron to reduce information loss and amplify global interaction representation. Finally, the computed attention weights are weighted to form the global normalized attention weights, which increases the utilization of effective information in input feature channels and spatial dimensions. We have optimized the backbone network, feature enhancement network, and detection heads to improve detection accuracy while ensuring a lightweight detection network. Specifically, we have added a small object detection layer to enhance the localization accuracy for the abundant small objects in aerial imagery. The algorithm's performance was evaluated using the publicly available VisDrone2019 dataset. Compared to the baseline network YOLOv8l, GNYL achieved a 7.2% improvement in $mAP_{0.5}$ and a 5.0% improvement in $mAP_{0.95}$. Compared to CDNet, GNYL showed a 14.5% improvement in $mAP_{0.5}$ and a 9.1% improvement in $mAP_{0.95}$. These experimental results demonstrate the strong practicality of the GNYL object detection network for detecting dense small objects in the aerial imagery captured by unmanned aerial vehicles.



Citation: Wu, H.; Zhu, Y.; Wang, L. A Dense Small Object Detection Algorithm Based on a Global Normalization Attention Mechanism. *Appl. Sci.* **2023**, *13*, 11760. <https://doi.org/10.3390/app132111760>

Academic Editor: Luis Javier Garcia Villalba

Received: 5 October 2023

Revised: 24 October 2023

Accepted: 26 October 2023

Published: 27 October 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/>).

Keywords: computer vision; small object detection; YOLOv8; attention mechanism; unmanned aerial vehicle

1. Introduction

Object detection algorithms and drone technology have become increasingly mature and play important roles in military and civilian domains, such as traffic management, agricultural irrigation, forest patrolling, and battlefield reconnaissance. These tasks can achieve higher accuracy and efficiency by using drones while also reducing casualties. As a result, aerial object detection has become a hot topic in computer vision. However, many problems still need to be solved for object detection in remote sensing images because of the wide image perspectives, multiple objects per image, and the large number of small objects [1]. Efficient feature extraction from the limited features of small objects in aerial images is a key to solving the above problems.

Among various object detection algorithms such as Faster R-CNN [2], SSD (Single Shot MultiBox Detector) [3], and RetinaNet [4], the YOLO (You Only Look Once) [5] object detection network has gained popularity due to its excellent performance in terms of both accuracy and detection speed. It has undergone network structure and module optimization in YOLOv3 [6], YOLOv5, and YOLOv7 [7]. And currently, YOLOv8 is the latest iteration that shows significant improvements in the model's light weight and detection speed.

The definition of small objects is divided into two categories: absolutely small objects with object pixels less than 32×32 in the COCO dataset and relatively small objects with object size less than 10% of the image size. To improve the detection performance of the YOLO object detection network for small objects, the original network is optimized and designed according to three main aspects: adding attention mechanism modules; incorporating multi-scale detection with feature enhancement networks; and dataset preprocessing [8,9].

The attention mechanism modules weigh the input features by mining more useful information in the channel or spatial dimensions, enhancing the perception of features along both the channel and spatial dimensions. In the channel attention mechanism proposed by JieHu et al., the SE (Squeeze and Excitation) [10] attention mechanism compresses and excites the channel information of the input features to obtain channel weights, which are then used to weigh the input features. However, the SE attention mechanism only considers features in the channel dimension and fails to capture spatial information, making it less suitable for dense small object detection. The CBAM (Convolutional Block Attention Module) [11] attention mechanism, proposed by Sanghyun Woo et al., combines channel and spatial attention modules in a serial manner to obtain attention weights, enhancing the utilization of spatial and channel information. However, the CBAM requires an additional fully connected layer and convolutional layer, resulting in a higher computational complexity and slower speed, requiring more hardware resources. The NAM (Normalization-Based Attention Module) [12] attention mechanism, proposed by Yichao Liu et al., obtains the attention weights by concatenating channel and spatial attention modules, utilizing information from different dimensional features and then improving the attention mechanism using weight factors. However, the NAM attention mechanism fails to fully explore the spatial information in the input features, resulting in lower utilization of spatial features. Some people have already integrated the interaction between the channel and spatial information into attention mechanisms, such as ECAP-YOLO [13], proposed by Munhyeong Kim et al., which considers the interaction between the channel and spatial information and has achieved positive results in small object detection. So, this paper proposed a GNAM (Global Normalization Attention Mechanism) considering the interaction between channel and spatial information, thereby preserving cross-dimensional information and utilizing the attention weights between channels, spatial width, and spatial height to improve efficiency.

Multi-scale detection in the feature enhancement network refers to inputting feature maps of different resolutions, which are obtained by fusing the detailed and semantic information of the original image into detection heads of multiple scales. This increases the richness of information inputted into the detection heads and the number and density of anchor boxes. Finally, the results from different scales are fused to obtain the final detection results. In response to the presence of multiple and dense objects in aerial images, the input features are at different scales from the downsampled data. Tsung YiLin et al. proposed an FPN (Feature Pyramid Network) [14], which extracts the features from the bottom-up backbone network and performs top-down upsampling, merging the features from the backbone network with the double-upsampled features to enhance the richness of details in the features. This improves the accuracy and efficiency of object detection. However, for dense small object detection, it is necessary to further increase the amount of output feature information. Therefore, we use a high-resolution feature enhancement network to improve the resolution of the network, which adds an upsampled layer into the FPN.

The overall idea of dataset preprocessing is to increase the number of small objects input into the network at the image level, thereby enhancing the objected weight of the trained detection network for small objects. The method of multiple-image stitching proposed by Chen Y et al. randomly selects four images with similar aspect ratios in the dataset and concatenates them into one image [15], which is then resized to a predefined size and inputted into the network. This method significantly improves the proportion of small objects in the input image. The Cutmix [16] method proposed by Yun S et al. randomly prunes the input image locally and fills it into a randomly selected local area

of the image in the dataset, thereby improving the network's detection performance for occluded objects.

Due to the challenges posed by dense small objects and a high proportion of small objects in the wide field of view of an aerial photography dataset, as well as the significant occlusion in certain scenes, leading to unclear object features and substantial differences between object detection box features in the dataset and pretraining datasets, it is challenging to achieve satisfactory results using generic object detection algorithms. Therefore, this paper proposes the integration of a GNAM (Global Normalization Attention Module) into the backbone network. The input features are processed using concatenated channel and spatial attention modules, preserving the cross-dimensional information. These modules suppress the irrelevant information in the input features while activating beneficial features for classification and localization tasks, enhancing the richness of detail and semantic information in the output features. They optimize the network structure for dense small objects by redesigning the backbone network, feature enhancement network, and detection head size. The goal is to increase the amount of feature information inputted into the detection heads while increasing the number and density of anchor boxes, thereby enhancing the network's perception capability for small objects. Based on the above two points, this paper proposes a Global Normalization Attention Mechanism-based object detection algorithm GNYL (Global Normalization Attention You Only Look Once).

In summary, the main contributions of this paper are as follows:

- (1) A new Global Normalization Attention Mechanism is proposed, which can suppress irrelevant information in the input features, enhancing the richness of detail and semantic information in the out features;
- (2) We demonstrate that GNYL can efficiently handle UAV images, enhancing robustness for small object detection;
- (3) We propose a new small object detection algorithm, GNYL, which can be applied more efficiently in practice.

2. GNYL (Global Normalization Attention Mechanism You Only Look Once)

This paper designs a GNYL object detection network based on YOLOv8l, with the network structure shown in Figure 1. The GNYL network structure consists of a backbone network and a detection head. The backbone network utilizes the CSPDarknet feature extraction network, while the head comprises a feature enhancement network and detection heads. This paper aims to improve the problem of dense small object detection in aerial datasets according to the following two aspects:

- (1) Adding the Global Normalization Attention Mechanism (GNAM): The GNAM first processes the input features using a channel attention unit, which utilizes scale factors from batch normalization (BN) to highlight features based on the variance measurement of the training model weights and then inputs them into a spatial attention unit, which incorporates spatial information using two convolutional layers, further emphasizing the spatial information. The interaction ability between the channel and spatial information is improved, fully exploits effective information for classification and localization tasks within the input features, and enhances the richness of detail in the GNAM's output features. In order to increase the detail richness of the GNAM's output features, the GNAM is added to the penultimate layer with rich semantic information in the backbone network;
- (2) The network architecture design, including the feature enhancement network and detection head design. Adding larger-scale detection heads increases the number and density of the anchor boxes, thereby improving the localization accuracy of small objects. A high-resolution feature enhancement network is employed to preserve more detailed information. Finally, utilizing large-scale detection heads with large and dense anchor boxes enhances the fitting between the predicted boxes and target boxes, thereby improving the localization accuracy.

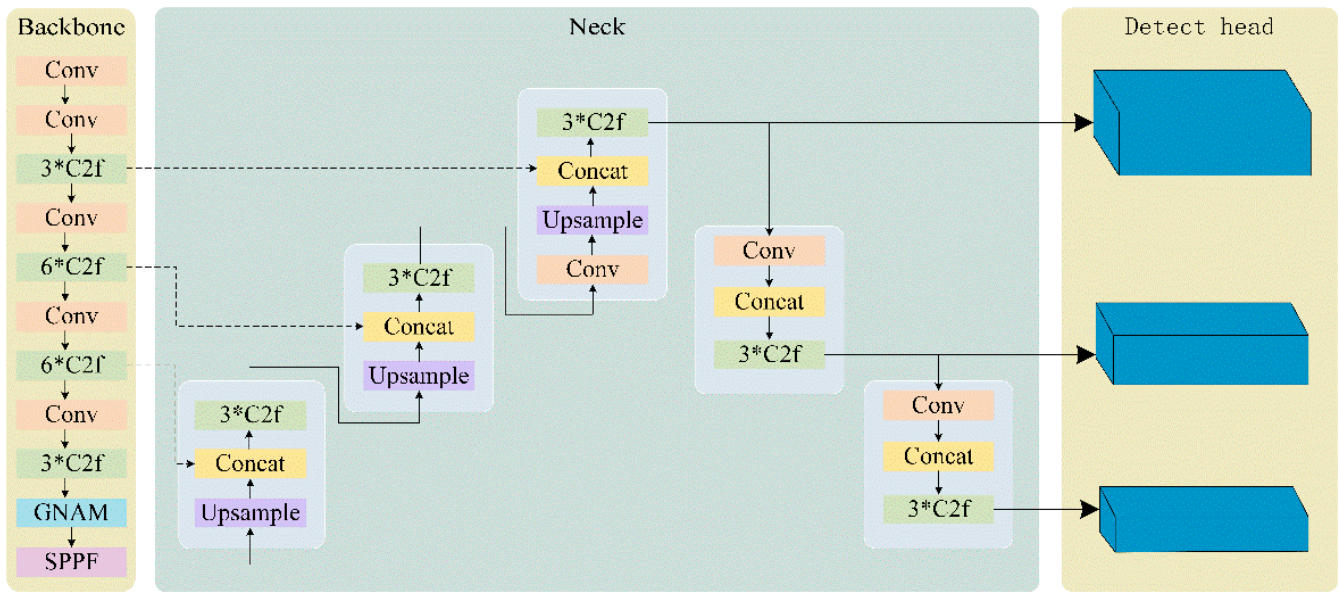


Figure 1. GNYL network structure diagram. “*” represents a multiplication sign.

2.1. Global Normalization Attention Mechanism

The design principle of the GNAM is to fully exploit the information in the input features while adding minimal network parameters and floating-point computations. This ensures the applicability of the plug-and-play GNAM attention module to different target detection networks and maintains the real-time performance of the object detection algorithm. The overall process of the GNAM is to input features using a channel attention unit, which utilizes scale factors from batch normalization (BN) to highlight features based on the variance measurement of the training model weights, and then inputs them into a spatial attention unit, which incorporates spatial information using two convolutional layers, further emphasizing the spatial information. The interaction ability between the channel and spatial information is improved, fully exploits effective information for classification and localization tasks within the input features, and enhances the richness of detail in the GNAM’s output features. The overall structure is shown in the Figure 2.

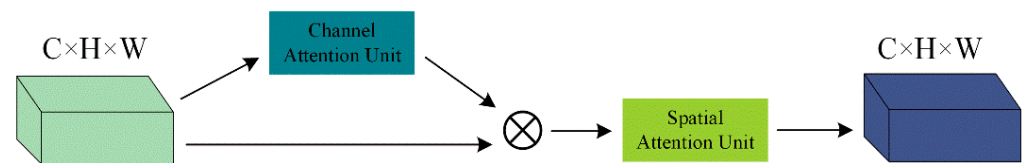


Figure 2. Global Normalization Attention Module structure diagram.

Assuming the input features X to the GNAM where $X = [x_1, x_2, \dots, x_c] \in \mathbb{R}^{C \times H \times W}$, they are passed through the channel attention unit, as shown in Figure 3. This paper adopts a scaling factor from batch normalization (BN), as shown in Equation (1). The scaling factor measures the variance of channels and indicates their importance.

$$B_{out} = BN(B_{in}) = \gamma \frac{B_{in} - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \quad (1)$$

where μ_B and σ_B are the mean and standard deviation of mini-batch B , respectively, and γ and β are the trainable affine transformation parameters (scale and shift) [17]. The channel attention unit is shown in Figure 3 and Equation (2). Batch normalization is performed on the input features, and the trainable weight applies to the results of batch normalization.

Then, the output feature M_c is obtained from the above results using the activation function.

$$M_c = F_{sigmoid}(W_\gamma(F_{BN}(X))) \quad (2)$$

where F_{BN} is batch normalization; γ is the scaling factor for each channel; the weights are obtained as $W_\gamma = \gamma_i / \sum_{j=0} \gamma_j$; and $F_{sigmoid}$ is the sigmoid activation function.

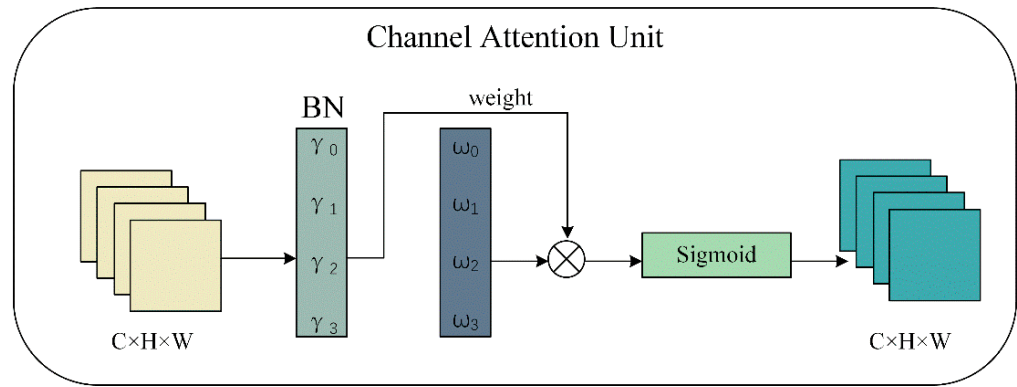


Figure 3. Channel attention unit.

In order to extract channel information while retaining more feature information, the output features of the channel attention unit are fused with the input features, achieving the goal of fully exploiting the effective information within the input features, as shown in Equation (3).

$$M = M_c \cdot X \quad (3)$$

In order to mine the spatial attention information, the input features M are passed through the spatial attention unit. The spatial attention unit without group convolution is shown in Figure 4. In this module, to concentrate on the spatial information, this paper adopts two 7×7 convolution layers for spatial information fusion. The 7×7 convolution layer has a larger receptive field that can cover a larger input feature region and provide more feature representations and better robustness, thus fully utilizing the spatial information of the features. This paper also adopts the same reduction ratio r as BAM [18]. Meanwhile, the max pooling layer significantly reduces the information, and contributes. This paper removes the pooling layer to preserve the features further. As a result, the spatial attention unit significantly increases the number of parameters. To prevent a notable increase in the parameters, this paper adopts group convolution with channel shuffling [19]; the output feature Y is obtained, as shown in Equation (6).

$$M_s = F_{ReLU}\left(F_{BN}\left(F_{Conv}^{7 \times 7}(M)\right)\right) \quad (4)$$

$$S = F_{Sigmoid}\left(F_{BN}\left(F_{Conv}^{7 \times 7}(M_s)\right)\right) \quad (5)$$

$$Y = F_{ChannelShuffle}(S) \quad (6)$$

where $F_{Conv}^{7 \times 7}$ represents the convolution layer using a convolution kernel of size 7×7 , F_{BN} is batch normalization, F_{ReLU} is the ReLU activation function, and $F_{ChannelShuffle}$ is the channel shuffling.

Embedding the plug-and-play GNAM into the backbone network of the object detection network can fully utilize the effective information of input features.

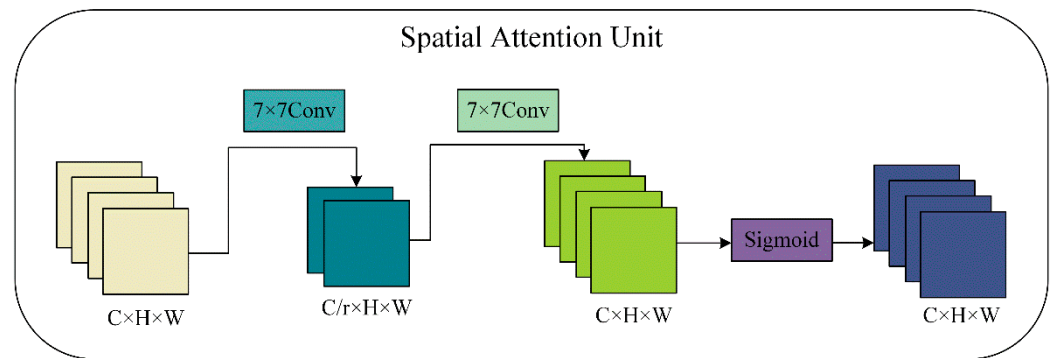


Figure 4. Spatial attention unit structure diagram.

2.2. High-Resolution Feature Enhancement Network

This paper mainly focuses on adapting high-resolution feature maps for dense small objects in aerial photography in the backbone network and detection head. Compared to the baseline network YOLOv8l, large-scale detection heads increase the density of the anchor boxes to improve object localization accuracy. A high-resolution feature enhancement network combined with large-scale detection heads was designed to enhance the detection ability of small objects in aerial photography, as shown in Figure 5. The high-resolution feature enhancement network undergoes two upsampling operations, expanding the feature size from 40×40 to 160×160 , which increases the network width and enhances the feature resolution. Considering the target box sizes in the dataset, the small-scale detection head with a size of 20×20 is eliminated, which has little impact on the coverage of the target boxes in the aerial dataset. Moreover, it improves the localization accuracy with a minimal increase in the number of parameters, as shown in Table 1.

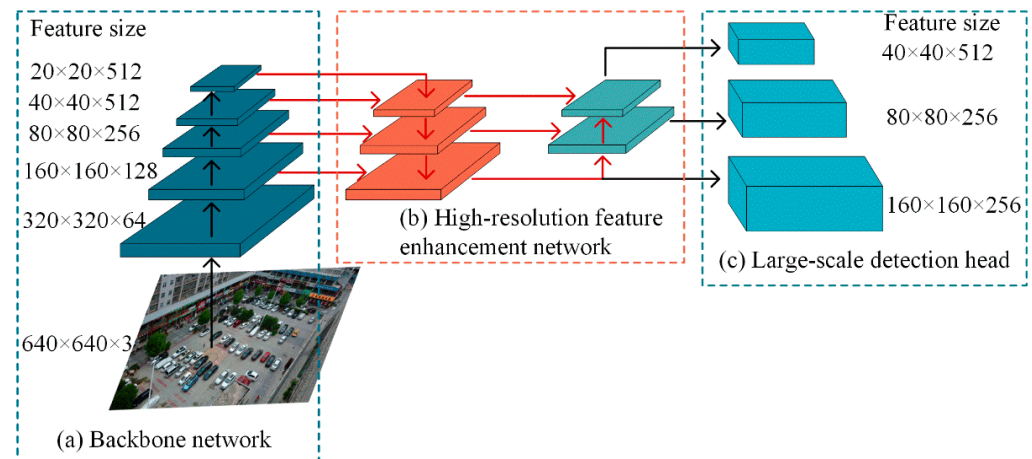


Figure 5. Network structure reconstruction diagram. (a) Backbone network; (b) high-resolution feature enhancement network; (c) large-scale detection head.

Table 1. Parameters for various network structures.

Model	Parameters	FLOPs	Head Size
YOLOv8l	43.6 M	164.9 G	20\40\80
YOLOv8l_rec	53.5 M	217.4 G	40\80\160

According to the distribution of the aspect ratios of objects and the sizes of the ground-truth boxes with the same center point in the VisDrone2019 training dataset (as shown in Figure 6), it can be observed that the aspect ratios of the objects are mainly distributed within 0.3 of the input image size. Additionally, there is a dense distribution of extremely small objects within 0.05 of the image size. This distribution pattern is consistent with the

analysis of the person and pedestrian categories in the dataset, indicating that they exhibit this characteristic.

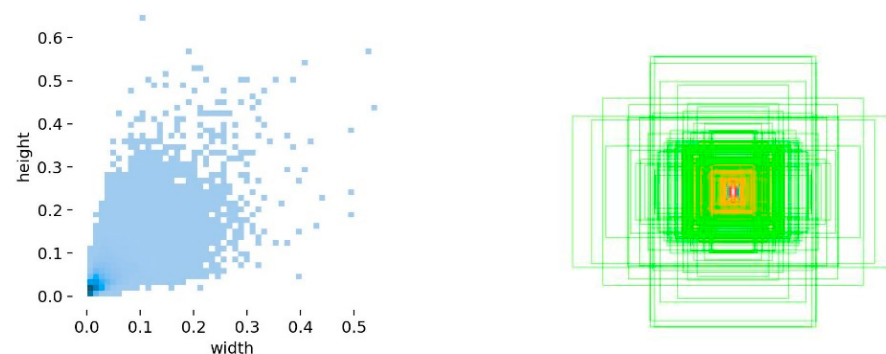


Figure 6. Training set object size feature plot.

Considering the target size and distribution characteristics of the object in the dataset, we enhance the number and density of the anchor boxes by introducing a large-scale detection head with a size of 160×160 . The object's center points are dispersed in different grids, and the initial anchor box size set in the large-scale detection head is closer to the size of the small object to be detected, thereby increasing the convergence speed. The predicted box obtained by adjusting the anchor boxes in different grids using training parameters has a higher degree of fitting with the ground-truth box, thereby improving the object detection accuracy.

3. Experiments

3.1. Datasets and Implementation Details

This paper uses the VisDrone2019 [20] public dataset, which consists of 6471 training images, 548 validation images, and 3190 test images (including 1580 images from VisDrone2019-DET-test-challen and 1610 images from VisDrone2019-DET-test-dev). All data splits are based on the original splits of each dataset. There is no data overlap between the training and testing sets. The dataset contains 10 classes of detection targets, namely pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. The challenges in object detection in the VisDrone2019 dataset are as follows:

- (1) Random changes in object size and shape;
- (2) The object is often obstructed by other objects, resulting in only partial object information being visible;
- (3) Images typically have large scales and high resolutions, requiring higher computational power;
- (4) Contains various types of objects and complex background environments.

In order to evaluate the effectiveness of GNYL, the following evaluation metrics are selected: parameter size, FLOPs (floating-point operations per second), and the mAP (mean average precision) at thresholds of 0.5 and 0.95. The experiments were conducted on an Ubuntu 18.04 system with an Intel(R) Xeon(R) Gold 6320R CPU @2.10 GHz, 128 GB RAM, NVIDIA GeForce RTX 3090 GPU, and PyTorch version 2.0.0. For the parameter settings during training, the image resolution for training models is 640×640 . We use mosaic data augmentation and mixup data augmentation; when using pre-trained weights, we freeze the backbone part for training for the first 10 epochs and then train the whole network; the learning rate is set to 0.01 and we use 90 training epochs.

3.2. Ablation Studies

To validate the effectiveness of the GNAM and the high-resolution feature enhancement network, an object detection network ablation experiment was conducted based on the YOLOv8l model.

Under the same conditions, YOLOv8l was selected as the baseline network for the comparative experiments to verify the impact of the GNAM and the high-resolution feature enhancement network with a large-scale detection head. Additionally, the SE, CBAM, GAM, CA, NAM and GNAM were added to the backbone network based on YOLOv8l, and the performance of different attention mechanisms was compared. The experimental results are shown in Table 2.

Table 2. Ablation experiments.

Method	mAP _{0.5}	mAP _{0.95}	Parameter	FLOPs
Baseline	41.6	25.0	43.6 M	164.9 G
YOLOv8l + GNAM	42.5	25.8	45.2 M	166.1 G
YOLOv8l + SE	41.5	25.1	43.7 M	165.5 G
YOLOv8l + CA	41.6	25.3	43.7 M	165.5 G
YOLOv8l + CBAM	41.6	25.2	43.6 M	165.4 G
YOLOv8l + GAM	41.8	25.2	48.8 M	194.1 G
YOLOv8l + NAM	42.1	25.5	48.2 M	189.9 G
YOLOv8l_rec	48.0	29.5	53.5 M	217.4 G
GNYL (Ours)	48.8	30.0	59.9 M	222.6 G

Using ablation experiments, it was concluded that compared to the baseline network YOLOv8l, GNYL achieved a 7.2% improvement in mAP at 0.5 and a 5.0% improvement in mAP at 0.95, while increasing the parameters by 37.3% and the floating-point computations by 34.9%.

When adding the SE, CA, CBAM, and GAM attention mechanisms to YOLOv8l, there was a slight increase in the parameters and floating-point computations, but there was almost no change in mAP_{0.5} and mAP_{0.95}. The addition of the SE, CA, CBAM, and GAM to the YOLOv8l network failed to effectively exploit the channel and spatial information of the input features. When adding NAM attention mechanisms to YOLOv8l, there was achieved a 0.5% improvement in mAP_{0.5} and mAP_{0.95}. Based on YOLOv8l_rec, GNYL incorporates the GNAM and uses batch normalization's scale factors to suppress irrelevant channels or pixels. It introduces a three-dimensional arrangement with a multi-layer perceptron for the spatial attention sub-module, reducing information loss and amplifying global interaction representation. Finally, the global normalization attention weights are weighted to increase the utilization of effective channel and spatial information in the input features.

The experimental results show that:

- (1) The scheme combining feature enhancement networks has significant accuracy advantages in detecting dense small objects in aerial photography;
- (2) The GNAM can fully mine the channel and spatial information of input features, effectively improving the utilization rate of input feature information.

3.3. Comparison of Detection Results of Different Object Detection Algorithms on VisDrone2019

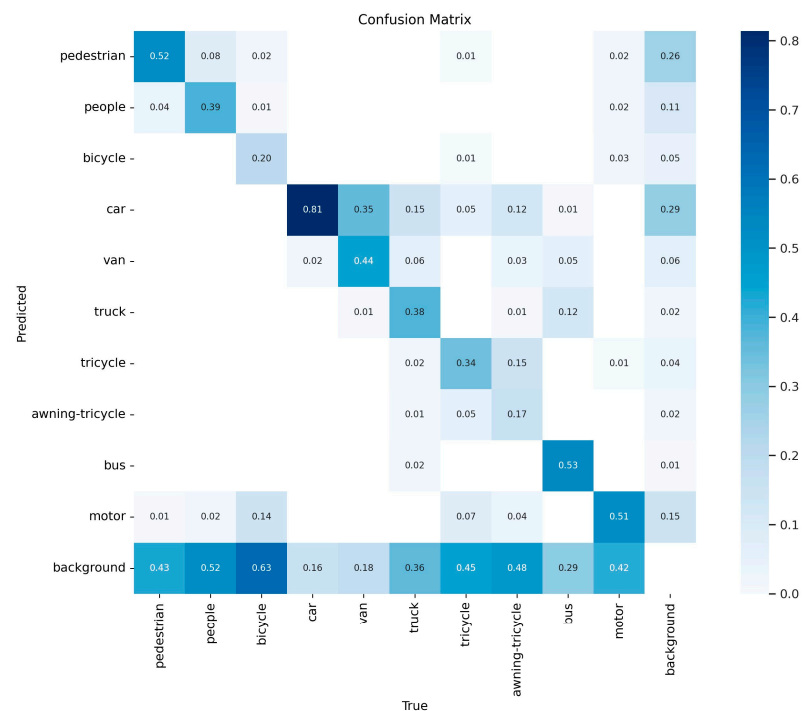
In order to validate the detection effect of GNYL on dense small objects in aerial images, the detection results of different object detection networks on VisDrone2019 were compared as shown in Table 3. The detection accuracy of pedestrian, people, bicycle, and motor objects showed a significant advantage with the highest numbers, with pedestrian detection increasing by 13.5% compared to the suboptimal network YOLOv5l's AP₅₀; There is also a significant advantage in detecting relatively large-object categories such as car, bus, and truck, with a mAP_{0.5} of 86.2% for car.

The comprehensive detection results of small and medium-sized objects in the large-sized categories and small-sized categories fully demonstrate that GNYL has significantly improved detection accuracy in object detection tasks aiming at dense small objects in aerial photography.

Table 3. Comparative experiments on different object detection algorithms.

Method	Backbone	Object Category										mAP _{0.5}
		Pedestrian	People	Bicycle	Car	Van	Truck	Tri	Awn-tri	Bus	Motor	
Faster R-CNN [21]	ResNet-50	21.4	15.6	6.7	51.7	29.5	19.0	13.1	7.7	31.4	20.7	21.7
Faster R-CNN [21]	ResNet-101	20.9	14.8	7.3	51.0	29.7	19.5	14.0	8.8	30.5	21.2	21.8
YOLOv4 [22]	CSPDarknet	24.8	12.6	8.6	64.3	22.4	22.7	11.4	7.6	44.3	21.7	30.7
CenterNet [23]	Hourglass-104	33.3	15.2	12.1	55.2	40.5	34.1	29.2	21.6	42.2	27.5	31.1
HR-Cascade++ [23]	HRNet-W40	32.6	17.3	11.1	54.7	42.4	35.3	32.7	24.1	46.5	28.2	32.5
CDNet [23]	ResNeXt-101	35.6	19.2	13.8	55.8	42.1	38.2	33.0	25.4	49.5	29.3	34.2
YOLOv5	CSPDarknet	44.4	36.7	18.5	74.2	37.7	37.4	25.3	12.7	48.6	43.3	37.9
GNYL (Ours)	CSPDarknet	57.9	46.3	22.0	86.2	53.1	42.3	37.6	21.0	64.5	57.4	48.8

To ensure the rationality of the experimental results, the algorithm is evaluated by referring to the widely used confusion matrix in existing object detection methods. The confusion matrix is the most basic, intuitive, and computationally simple method for measuring the accuracy of classification models. The horizontal axis represents the real label, and the vertical axis represents the predicted results of the model. The confusion matrix of GNYL on VisDrone2019 is shown in Figure 7.

**Figure 7.** The confusion matrix of GNYL on VisDrone2019.

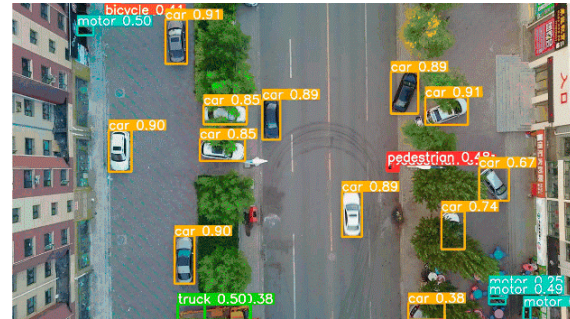
We evaluate all predictions for the test set of VisDrone2019. The object detection confusion matrix mainly focuses on the detection results of the detection frame rather than the entire image. The diagonal represents the probability that the model detects a certain class of Visdrone2019 data, and its true label is also a certain class, while the remaining areas represent false or missed detections. Our model performs well in categories with high background contrast, such as car, bus, motor, and pedestrian, especially in car, where the true positive is as high as 81%. In categories with low contrast with the background, such as bicycle, tricycle and awning-tricycle, the performance is average, and some categories are predicted as background information. In addition, our model has a low error detection rate and rarely predicts one category as being another. From this, it can be seen that our model exhibits high accuracy in common object categories and can be well applied on unmanned aerial vehicles.

In order to validate the effectiveness of GNYL in different scene detection scenarios, detection samples were selected from the VisDrone2019 test set based on light changes, high-

altitude field of view, complex backgrounds, and object occlusion, as shown in Figure 8. At the same time, the detection performance of the baseline network and GNYL are compared under the limit of selecting a rich number of small objects, as shown in Figure 9.



(a) Detection image under changes in light



(b) Detection image under high-altitude field of view



(c) Detection image under complex background

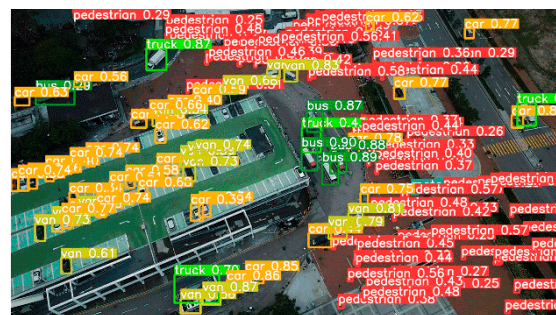


(d) Detection image under target occlusion

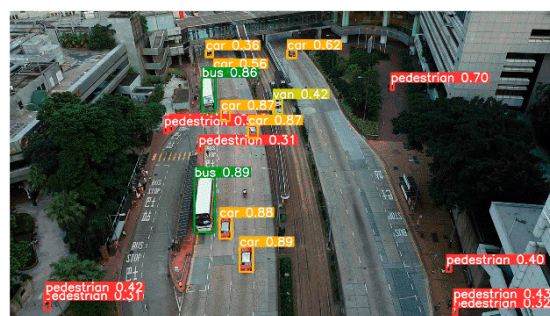
Figure 8. GNYL detection effect in different scenarios.



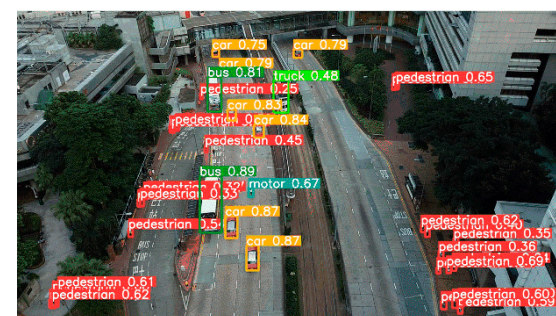
(a) Example 1 (Baseline)



(b) Example 1 (GNYL)



(c) Example 2 (Baseline)



(d) Example 2 (GNYL)

Figure 9. Comparison of detection effects.

4. Conclusions

This paper proposes the GNYL (Global Normalization Attention Mechanism You Only Look Once) object detection network based on the YOLOv8l detection network, specifically targeting the task of detecting dense small objects in aerial images. Firstly, the GNAM is proposed, which processes the input features using a channel attention unit, which utilizes scale factors from batch normalization (BN) to highlight features based on the variance measurement of the training model weights, and then inputs them into a spatial attention unit, which incorporates spatial information using two convolutional layers, further emphasizing the spatial information. The interaction ability between the channel and spatial information is improved, fully exploits effective information for classification and localization tasks within the input features, and enhances the richness of detail in the GNAM's output features. In order to increase the richness of detail of GNAM's output features, the backbone network and feature enhancement network are redesigned to address the challenges posed by small objects in aerial images. This involves increasing the amount of feature information inputted into the detection heads and adjusting the number and density of anchor boxes. These modifications aim to improve the accuracy of small object detection.

The performance of the proposed GNYL network is evaluated using the publicly available aerial dataset VisDrone2019. The GNYL network significantly improves the detection accuracy compared to the baseline network. GNYL achieves the highest mean average precision (mAP) for 9 of 10 object categories in the VisDrone2019 dataset among various object detection networks. Experiments have shown that GNYL has strong practicality in dense small object detection tasks.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

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

References

1. Jiang, B.; Qu, R.K.; Li, Y.D.; Li, C. Object detection in UAV imagery based on deep learning: Review. *Acta Aeronaut. Astronaut. Sin.* **2021**, *42*, 137–151.
2. Ren, S.; He, K.; Girshick, R.; 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](#)]
3. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot MultiBox detector. In *Proceedings, Part I 14, Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
4. 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.
5. 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.
6. Redmon, J.; Farhadi, A. Yolo3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
7. 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, CO, Canada, 18–22 June 2023*; pp. 7464–7475.
8. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [[CrossRef](#)]

9. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [[CrossRef](#)]
10. 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.
11. Sanghyun, W.; Jongchan, P.; Joon-Young, L.; In, S.K. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
12. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based attention module. *arXiv* **2021**, arXiv:2111.12419.
13. Kim, M.; Jeong, J.; Kim, S. ECAP-YOLO: Efficient channel attention pyramid YOLO for small object detection in aerial image. *Remote Sens.* **2021**, *13*, 4851. [[CrossRef](#)]
14. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
15. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Meng, G.; Jia, J. Stitcher: Feedback-driven data provider for object detection. *arXiv* **2020**, arXiv:2004.12432, 12.
16. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
17. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
18. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
19. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
20. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, L.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
21. 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, Virtual, 5–9 January 2021; pp. 3258–3267.
22. Ali, S.; Siddique, A.; Ateş, H.F.; Güntürk, B.K. Improved YOLOv4 for aerial object detection. In Proceedings of the 29th Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey, 9–11 June 2021; pp. 1–4.
23. Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, D.; Zhang, J.; Zhu, P.; Liu, M.; et al. VisDrone-DET2021: The vision meets drone object detection challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2847–2854.

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.