Application Research of Fast UAV Aerial Photography Object Detection and Recognition Based on Improved YOLOv3

Yan Xia\textsuperscript{1a}, Guoxiang Ye\textsuperscript{1b}, Songsong Yan\textsuperscript{2c}, Zangwei Feng\textsuperscript{1}, Feng Tian\textsuperscript{1}

\textsuperscript{1}Technology Center, Chengdu Aircraft Industry (Group) Co., Ltd., Chengdu, China
\textsuperscript{2}Anhui Kaidian Automation Equipment Co., Ltd., Wuhu, China
\textsuperscript{a}yanxiahit@163.com, \textsuperscript{b}ygyiyec@hotmail.com, \textsuperscript{c}xiaosongshine@163.com

Abstract: UAV aerial photography is affected by weather, altitude, illumination, occlusion and other factors resulting in the objective fact that the object has many variations in scale and perspective. This paper proposes an improved YOLOv3 algorithm which achieve object detection and recognition in view of the above complex working conditions quickly and accurately. Firstly, the network framework of YOLOv3 is altered and the BN layer is integrated into the convolution layer. The network structure is simplified and the speed of model detection is greatly accelerated but the detection accuracy is almost the same as original. Secondly, GIOU is adopted as the loss function for bounding box regression to prevent the existence of zero gradient and improve mean average precision (mAP) of the detection model. Finally, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes independently. The results of a large number of comparative experiments show that the improved YOLOv3 algorithm can achieve object detection for the UAV aerial photography faster and more accurately. The detection speed is increased by 27\%, the mAP is increased by 2.32\% and the detection performance is also improved for small object significantly compared with YOLOv3.

1. Introduction
UAV aerial photography plays a major role in military intelligence processing at present. The rapid aerial photography object detection and recognition has great advantages in the future war. The UAV equipped with industrial camera is applied to take aerial photos to obtain intelligence data in different levels of airspace. UAV aerial photography has the advantages of low cost, strong safety performance and high stability compared with UAV operation. It has been widely applied in military terrain reconnaissance, intelligence reconnaissance, road survey and electric power inspection. The object data has the diversity of scale and the particularity of viewing angle with the height and angle of UAV aerial photography changing constantly. The size of the object data from aerial photography may be only a dozen or even several pixels when the aerial photography object is far away from the UAV, which poses a greater challenge to the object detection and recognition algorithm, as shown in Fig.1(a). At the same time, aerial photography will also be affected by complex environmental background such as cloud occlusion, light intensity change and time variation, which puts forward higher requirements for object detection and recognition algorithm, as shown in Fig.1(b).
Deep learning has been widely applied in object detection field. At present, there are three main categories of object detection and recognition methods in industry and academia. Namely, traditional detection algorithm, one-stage detection algorithm and two-stage detection algorithm. General traditional detection algorithms include DPM\textsuperscript{[1]}, HOG/ SVM\textsuperscript{[2]}, Haar / SVM\textsuperscript{[3]} and many optimizations of the above methods. These traditional detection algorithms are based on the region selection strategy of sliding window or matching according to the feature points. Therefore, they have weak robustness, high time complexity and window redundancy to the variety of object diversity. The main methods are R-CNN\textsuperscript{[4]}, SPP net\textsuperscript{[5]}, Fast RCNN\textsuperscript{[6]}, Faster RCNN\textsuperscript{[7]}, Cascade RCNN\textsuperscript{[8]} and R-FCN\textsuperscript{[9]} for the two-stage detection algorithm. Two-stage method is equivalent to training two models, which will lead to the increase of training parameters and training time and affect the detection efficiency. Based on the one-stage detection algorithm includes the YOLO series, SSD\textsuperscript{[10]}, Retinanet\textsuperscript{[11]} mainly. They abandon the disadvantages of RCNN based preselected box, such as large memory consumption, low calculation speed and repeated feature extraction. The idea of end-to-end fast detection is realized without affecting the accuracy of object detection.

At present, neither one-stage detection algorithm nor the two-stage detection algorithm has a good effect on small object detection. At the same time, although the detection accuracy of the one-stage algorithm can be compared with that of the two-stage algorithm, it still needs to be improved in the industry. In this paper, YOLOv3 algorithm is improved in order to improve the precision and speed of object detection and identify small objects accurately. The main contributions of this paper are as follows:

1. BN layer parameters are integrated into the convolution layer in the model, which reduces the training parameters and the amount of calculation. Finally, shorten the object detection time of UAV aerial photography greatly but the accuracy is almost the same as the original.

2. Aiming at the YOLOv3 loss function, the boundary box regression loss is improved and GIOU is applied as loss function for the boundary box regression to improve the mAP of UAV aerial photography object detection.

3. A new scale is designed for fusion detection and a pyramid with four different scale convolutions is constructed together with the original three scales of YOLOv3 and the resolution of the feature map.
is 160 * 160, 80 * 80, 40 * 40, 20 * 20. In order to improve the small object detection accuracy in UAV aerial images, four different scale feature pyramids are fused by the up-sampling operation.

2. Related work

YOLO\textsuperscript{[12]} has been widely applied in industry as a representative of the two-stage detection algorithm. In 2015, Redmon proposed the YOLO detection algorithm, which improved the detection speed compared with the two-stage detection algorithm greatly. YOLOv2 \textsuperscript{[13]} optimized the YOLO network structure and reduced the complexity of model training by using the anchor frame mechanism. K-means cluster analysis method is applied to determine the number of anchor frames, which increases model prediction accuracy. YOLOv3 \textsuperscript{[14]} optimizes the network model on the basis of YOLOv2. There is no pooling layer and full connection layer in the whole YOLOv3 network. Multi-scale fusion method is applied to detect and recognize different scales objects with reference to the FPN network structure.

In view of different detection objects, domestic and overseas scholars have proposed many improved YOLOv3 methods. Paper\textsuperscript{[15]} proposes to enhance the residual network of YOLOv3. Six different scales convolution feature maps are designed for the larger objects in the image and fused with the corresponding scale feature map in the residual network to prevent the occurrence of object detection error in complex scenes. This paper\textsuperscript{[16]} proposed a slimYOLOV3 with stronger real-time performance that can be deployed to the UAV. SlimYOLOV3 is superior to the original YOLOv3 algorithm in terms of parameter number, memory use, inference time and other aspects. Stanford scholars\textsuperscript{[17-18]} proposed that it is the best choice to optimize with the metric itself as the loss but IOU is zero and its gradient is also zero in the bounding box non overlapping scene, so it affects the training quality and convergence rate. Therefore, GIOU will always have an effective gradient, which can guide how to optimize the model continuously and effectively as the loss function of bounding box regression. The papers\textsuperscript{[19-20]} only modified the network structure to some extent and improved the detection speed by reducing the existing YOLOv3 network architecture but the precision decreased. In order to solve the problem of detecting small objects effectively. The paper\textsuperscript{[21-23]} changed the shape and quantity of anchor and increased the feature fusion object detection layer after down-sampling to detect small objects, which have strong robustness for small scale and occluded objects.

3. Improved YOLOv3 Algorithm

3.1 Fast UAV Aerial Object Detection

The training model usually adds BN layer after convolution layer in the task of object detection. The BN layer can solve the problem of gradient disappearance and explosion after normalizing the data effectively. At the same time, BN layer can speed up the network convergence speed and improve the network generalization ability. The activation input value of deep neural network will shift with the deepening of the network depth before the nonlinear transformation. Adding BN layer is to force the more and more biased distribution back to the standard distribution, so that the activated input value can fall in the area where the nonlinear function is sensitive to the input, so as to speed up the convergence process. The parameter-adjusted process is relatively simple and low requirements for initialization parameters with the BN layer.

Although the BN layer plays a positive role, it increases the network layer operations in the process of network forward inference, which affects the model performance and takes up a lot of graphics memory space. At present, many advanced networks use the BN layer, so it is necessary to integrate the parameters of BN layer into the convolution layer, which can accelerate the forward inference speed of the training model enormously. The specific fusion process is as follows:

\begin{align}
X_i &= W_{\text{conv}} \ast x_i + b_{\text{conv}} \\
Y_i &= \frac{X_i - E(X_i)}{\sqrt{\text{Var}[X_i] + \epsilon}} \ast \gamma + \beta
\end{align}
Among them, convolution weight $W_{conv} \in \mathbb{R}^{C \times C}$, convolution bias $b_{conv} \in \mathbb{R}^C$, are used to calculate the mean and variance of the model in each mini batch instead of the population mean and variance but the population mean and variance cannot be calculated in the prediction, so the mean and variance of the batch in the training is used to estimate them. The final formula (3) is the result of fusing the convolution layer and BN layer of the model. The new convolution weight and bias terms are as follows:

$$W_{new_{conv}} = \frac{W_{conv} \cdot \gamma}{\sqrt{\text{Var}[W_{conv} \cdot x_i + b_{conv}]} + \varepsilon}$$

$$b_{new_{conv}} = \left( \frac{b_{conv} - E(W_{conv} \cdot x_i + b_{conv})}{\sqrt{\text{Var}[W_{conv} \cdot x_i + b_{conv}]} + \varepsilon} \right) \cdot \gamma + \beta$$

BN layer can be deleted and convolution operation with new weight and bias through the above calculation, so as to obtain the same batch normalization effect as the original.

### 3.2 GIOU as Loss Function for Bounding Box Regression

YOLOv3 uses MSE as the loss function for bounding box regression in the object detection task. The actual MSE loss function is quite sensitive to the scale invariant of the object. Even though YOLOv3 reduces the influence of scale invariant on the regression accuracy by the means of taking a square to length and width, the effect is not obvious. However, it will be improved under certain circumstances if the IOU is used as the loss function for bounding box regression. The IOU will be zero if there is no coincidence between the detection box and the bounding box. Obviously, the gradient will be zero and cannot be optimized when IOU is used as loss for bounding box regression. At the same time, the detection effect will be greatly different when there is the same IOU between the detection box and the bounding box because the IOU cannot distinguish the different alignment between two objects. They will have exactly the same IOU if there are two overlapping objects with the same intersection set in different directions.

GIOU\textsuperscript{[18]} will have a good detection effect as loss function for bounding box regression of object detection. GIOU has excellent properties such as non negativity, identity and symmetry as a metric index. It also has the scale invariant of IOU. GIOU can still calculate the gradient when there is no intersection between the detection box and the bounding box. The gradient will not be zero and continue to optimize the model. GIOU overcomes the disadvantage of IOU as the loss function to a certain extent. As loss function for bounding box regression, GIOU is as follows:

$$\text{IOU} = \frac{|A \cap B|}{|A \cup B|}$$

$$\text{GIOU} = \text{IOU} - \frac{|C - A \cup B|}{|C|}$$

$$L_{\text{GIOU}} = 1 - \text{GIOU}$$

Any two convex shapes $A, B \subseteq S \in \mathbb{R}^n$, find the smallest closed convex object $C$ for $A, B$, $C \subseteq S \in \mathbb{R}^n$. 
3.3 Multiscale Feature Fusion Object Detection
In general, the feature semantic information of the bottom layer is less but the object location information is accurate. The feature semantic information of the top layer is rich but the object location information is fuzzy. YOLOv3 adopts the method similar to up-sampling and fusion on FPN. Three different scales are merged and fused to detect the object independently on the multi-scale fusion feature map.

As the aerial photograph changes with the altitude, the object will become smaller and smaller and different weather conditions will also affect the small objects imaging to a certain extent. Darknet-53 is applied as backbone infrastructure network for object detection feature extraction according to the characteristics of UAV aerial photograph objects and four different convolution layers are used for feature extraction with resolutions of 160 * 160, 80 * 80, 40 * 40, 20 * 20. The up-sampling operation is applied to fuse with four different scales feature pyramids to detect and identify the object independently in each branch and obtains the deep fusion detection model finally, as shown in Fig.2.

![Fig.2 Fast detection UAV aerial photograph model structure](image)

4. Experiment and Analysis
In the training stage, the momentum parameter is 0.9, the weight attenuation regular term parameter is 5e-4 and the mini-batch random gradient descent method is applied to learn the model. The initial learning rate parameter is 1e-3 and the learning rate is adjusted to 5e-4 when the training reaches 2000 batches. The learning rate is adjusted to 5e-5 when the training reaches 4000 batches. The training data is enhanced by rotating the image at different angles, adjusting the saturation, adjusting the exposure and adjusting the hue.

In order to verify that the improved YOLOv3 algorithm can detect and recognize the UAV aerial photograph quickly. This paper uses the experiment of comparing with the traditional YOLOv3 to train in the UAV aerial photograph data set. Through the comparative analysis a large number of experimental results to demonstrate whether the improved YOLOv3 algorithm can quickly and accurately identify the object in the UAV aerial photograph. All experiments are based on the Darknet[24] framework and run on a PC equipped with Intel Core i9-9900k CPU and NVIDIA 1080ti GPU.

4.1 Data Set Description
The data set of this paper comes from the first aviation Cup object detection and recognition competition. There are 4356 UAV aerial images in different weather, height, light and urban areas. The maximum number of UAV aerial photograph object to be detected is more than 1000 with different degrees occlusion and truncation in each image. The data set can verify the detection model whether it has strong robustness. 3233 of them are randomly selected as the training set and 1123 as the test set. Fig.3 shows the overall working condition distribution of the UAV aerial photograph data set.
4.2 Fusion BN Layer Experimental Analysis

In order to improve the detection speed of the detection model, the BN layer is fused into the convolution layer in the network structure. As shown in Table 1, it can be seen that for 1123 test set pictures through the comparative experimental analysis, the total detection time of the network model without fusion is 82.57s, while the detection time of the BN layer fused into the convolution layer is 64.97s and the speed is increased by 27% but the mAP is reduced by 0.31%. Therefore, the fusion technology can enormously accelerate the object detection speed of UAV aerial photograph object while the mAP is basically unchanged, which has won precious time for the post-processing of UAV aerial photograph.

| Network Structure      | Input Size | Detection Time (s) | mAP@0.5% |
|------------------------|------------|--------------------|----------|
| Convolution fusion BN  | 640*640    | 64.97              | 92.15    |
| Convolution unfused BN | 640*640    | 82.57              | 92.43    |

4.3 Loss Function and Metric Index Influence on Model Performance

This paper uses the improved evaluation index of the 2018 mscoco challenge\textsuperscript{[25]} to calculate the test results. The evaluation indicators of the 2018 mscoco challenge include the calculation of mAP and determine true positives and false positive of test samples by setting different IOU thresholds. The main performance metric index is displayed by mAP which is composed of different IOU thresholds, i.e. IOU = \{0.5, 0.55, 0.95\}. The mAP with different thresholds is calculated respectively and the final evaluation index is obtained by weighted average calculation. It is meaningless to increase the IOU threshold to 0.9 in the actual working condition because of the particularity of UAV aerial photograph data. So the IOU threshold is only set as IOU = \{0.5, 0.55, ..., 0.8\}. At the same time, we also modify the metric index and apply GIOU metric index to judge the true positive and false positive and form comparative experiments with IOU. This paper also lists IOU and GIOU threshold as 0.65 separately which is shown as mAP.65 in the table.
Table 2: Different loss functions result for different metrics on test

| loss /Evaluation | mAP@.50,55,60,65,70,75,80% | mAP@.65% |
|------------------|----------------------------|-----------|
|                  | IOU | GIOU | IOU | GIOU |
| MSE              | 77.42 | 75.16 | 85.01 | 83.94 |
| LIoU             | 78.72 | 77.96 | 86.38 | 85.58 |
| Relative improv.%| 1.68% | 3.73% | 1.61% | 1.95% |
| LIOU             | 80.12 | 79.43 | 88.43 | 87.69 |
| Relative improv.%| 3.48% | 5.68% | 4.02% | 4.47% |

As shown in Table 2, it can be seen that setting the loss function for bounding box regression LIOU and LIoU has a certain improvement effect under the same metric index. Especially, the effect is the best when the measurement index is GIOU and the loss function LIOU is used, the map is increased by 5.68%. From formula (5) and formula (6), it can be concluded that the actual coincidence degree of the bounding box and the prediction box is different when the metric index GIOU and IOU are equal, the corresponding coincidence degree of GIOU is less than IOU. It can also be seen from Table 2 that the evaluation index map using GIOU as the metric index is slightly less than IOU when the threshold value of IOU and GIOU is the same, which is more obvious when the loss function is MSE Obvious. On the whole, LIOU is loss function for the bounding box regression plays an important role in the detection model precision for the UAV aerial photograph data set. The detection effect is shown in Fig. 4.

Variation for Object occlusion

Variation for UAV aerial photograph height

Variation for light intensity

Fig. 4 Results for object detection and recognition for different working conditions
4.4 Multi-scale Fusion Experimental analysis
It is extremely significant for information processing to be able to recognize and detect small objects or distant objects in high-resolution scene images. For many objects, such as traffic signs or cars, can hardly be seen in high-resolution images which poses a greater challenge for the detection model. Small object detection can be solved by increasing the resolution of input image or by fusing high-resolution features and high-dimensional features of low-resolution image. The original multi-scale fusion strategy of YOLOv3 is to detect objects on three different scale feature maps. The scale of these three feature maps is 13 * 13, 26 * 26, 52 * 52 when the input is 416 * 416. But small object data accounts for a large proportion for UAV aerial photograph data and the size of training object data is not uniform, mainly 1920 * 1080, 4096 * 2160, 3840 * 1641. There are hundreds of objects to be detected in an aerial photograph data which has formed a huge test for the detection model. Increase input the resolution of the image in the case of using the loss function for bounding box regression in 4.3 and apply three different resolutions for the model input to carry out the contrast analysis training. Four different scales are integrated to detect the object independently on the feature map in addition to the improvement of the loss function, as shown in table 3.

| network structure | YOLO layer | Characteristic pyramid size | mAP@.5% |
|-------------------|------------|----------------------------|---------|
| YOLOV3-3          | 3          | 80*80,40*40,20*20           | 92.43   |
| YOLOV3-4          | 4          | 160*160,80*80,40*40,20*20   | 94.57   |

K-means clustering method is applied to automatically generate for the selection of anchor frames. Four convolution feature maps of different scales are used to detect and identify objects on the UAV aerial photograph data. Each scale corresponds to three anchor boxes with a total of 12 anchor frames, (33 * 50), (46 * 91),(53 * 28),(58 * 88), (76 * 139),(78 * 93),(84 * 66),(89 * 43),(95 * 54),(100 * 74),(139*67),(190*108). The final result of small object detection and recognition of UAV aerial photograph is shown in Figure 4-4, about 1000 objects.

Fig. 5 Results for small object detection and recognition

5. Conclusion
In this paper, a method of object detection based on improved YOLOv3 UAV aerial photography is
proposed. It is demonstrated that the fusion BN layer can significantly improve the speed of object detection. GIOU and IOU are applied as the metric indexes of detection model respectively and three kinds of loss functions for bounding box regression, MSE, LIOU and LGIOU are compared and analyzed. The mAP of object detection is significantly improved in the UAV aerial photography data set. The detection model based on multi-scale feature fusion can also improve the detection performance of small objects. In order to detect more detailed information content, it is necessary to identify the UAV aerial photography object detection with fine-grained in the follow-up work including the object color, multi-attribute category and explore the more robust object detection model.

### Authors’ background

| Your Name          | Title*                  | Research Field                        | Personal website                          |
|--------------------|-------------------------|---------------------------------------|-------------------------------------------|
| Yan Xia            | Engineer                | UAV control, image processing         | https://orcid.org/0000-0001-7855-8192    |
| Guoxiang Ye        | Engineer                | UAV intelligence processing           |                                           |
| Songsong Yan       | Engineer                | Image processing                      |                                           |
| Zhangwei Feng      | Engineer                | UAV intelligence processing           |                                           |
| Feng Tian          | Researcher, Engineer    | UAV control, image processing         |                                           |

### Reference

[1] Jonathan B. Freund, Jonathan F. MacArt. DPM: A deep learning PDE augmentation method (with application to large-eddy simulation). ArXiv:1911.09145, 2019.

[2] Gustavo Neves, Rômulo Cerqueira, Jan Albiez, Luciano Oliveira. Rotation-invariant shipwreck recognition with forward-looking sonar. ArXiv:1910.05374, 2019.

[3] Yu Guang Wang, Ming Li. HaarPooling: Graph Pooling with Compressive Haar Basis. ArXiv:1909.11580, 2019.

[4] Girshick R, Donahue J, Darrell T. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR, 2014, 580-587.

[5] He K M, Zhang X Y, Ren S Q. Spatial pyramid pooling in deep convolutional networks for visual recognition. ArXiv:1406.4729v4, 2015.

[6] Girshick R. Fast R-CNN. ICCV, 2015: 1440-1448.

[7] Ren S Q, He K M, Girshick R. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,39(6), 1137-1149.

[8] Z Cai, N Vasconcelos. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. ArXiv:1906.09756, 2019.

[9] Dai JF, He KM, Sun J. R-FCN: object detection via region-based fully convolutional networks[J]. ArXiv:1605.06409, 2016.

[10] Liu W, Anguelov D, Erhan D. SSD: single shot multibox detector. ECCV, 2016, 21-37.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He. Focal loss for dense object detection. The IEEE International Conference on Computer Vision (ICCV), 2017, 2980-2988.

[12] Redmon J, Divvala S, Girshick R. You only look once: Unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition, 2016, 779-788.

[13] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger. IEEE Conference on Computer Vision and Pattern Recognition, 2017, 6517-6525.

[14] Redmon J, Farhadi A. YOLOv3: An incremental improvement. IEEE Conference Computer Vision and Pattern Recognition, 2018.

[15] ZHANG Fukai, YANG Feng, LI Ce. Fast vehicle detection method based on improved YOLOv3. Computer Engineering and Applications, 2019.

[16] Pengyi Zhang, Yuxin Zhong, Xiaojiong Li. SlimYOLOv3: Narrower, Faster and Better for Real-Time UAV Applications. The IEEE International Conference on Computer Vision (ICCV),
2019.

[17] Z Zheng, P Wang, W Liu, J Li, R Ye. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. ArXiv, 2019.

[18] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 658-666.

[19] Rui Huang, Jinan Gu, Xiaohong Sun. A Rapid Recognition Method for Electronic Components Based on the Improved YOLO-V3 Network Electronics 2019, 8(8), 825.

[20] Ji-Hun Won, Dong-Hyun Lee, Kyung-Min Lee. An Improved YOLOv3-based Neural Network for De-identification Technology. ITC-CSCC, 2019, 3(6), 117.

[21] Moran Ju, Haibo Luo, Zhongbo Wang. The Application of Improved YOLO V3 in Multi-Scale Object Detection. Appl. Sci, 2019, 9(18), 3775.

[22] KJ Kim, PK Kim, YS Chung. Performance Enhancement of YOLOv3 by Adding Prediction Layers with Spatial Pyramid Pooling for Vehicle Detection. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, 8(6), 394.

[23] by Jing Li, Jinan Gu, Zedong Huang, Jia Wen. Application Research of Improved YOLO V3 Algorithm in PCB Electronic Component Detection. Appl. Sci, 2019, 9(18), 3750.

[24] Redmon J. Darknet: open source neural networks. http://pjreddie.com/darknet/.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona. Microsoft coco: Common objects in context. In European conference on computer vision, Springer, 2014. 2, 5, 7.