# YOLOv5-Based Dense Small Target Detection Algorithm for Aerial Images Using DIOU-NMS

*Yu WANG, Xiang ZOU, Jiantong SHI, Minhua LIU**

School of Artificial Intelligence, Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, 100048, Beijing, China

wangyu@btbu.edu.cn, 571529288@qq.com, zouxiang523@126.com, 2015952331@qq.com*

Submitted September 8, 2023 / Accepted November 7, 2023 / Online first December 9, 2023

**Abstract.** *With the advancement of various aerial platforms, there is an increasing abundance of aerial images captured in various environments. However, the detection of densely packed small objects within complex backgrounds remains a challenge. To address the task of detecting multiple small objects, a multi-object detection algorithm based on Distance Intersection Over Union loss Non-Maximum Suppression (DIOU-NMS) integrated with You Only Look Once version 5 (YOLOv5) is proposed. Leveraging the YOLOv5s model as the foundation, the algorithm specifically addresses the detection of abundantly and densely packed targets by incorporating a dedicated small object detection layer within the network architecture, thus effectively enhancing the detection capability for small targets using an additional upsampling operation. Moreover, conventional non-maximum suppression is replaced with DIOU-based non-maximum suppression to alleviate the issue of missed detections caused by target density. Experimental results demonstrate the effectiveness of the proposed method in significantly improving the detection performance of dense small targets in complex backgrounds.*

## Keywords

Object detection, YOLOv5, DIOU-NMS, aerial images, small object detection, complex backgrounds

## 1. Introduction

Object detection has widespread applications in various fields, including autonomous driving, intelligent transportation, security surveillance, etc. In real-world scenarios, aerial images often contain a multitude of dense small targets which may undergo deformations, occlusions, and susceptibility to adverse effects from complex backgrounds characterized by varying lighting and clutter interference. Consequently, motion object detection encounters significant challenges in such scenarios.

Object detection is a vital field in computer vision, integrating cutting-edge technologies from image processing, pattern recognition, automatic control, artificial intelligence, and computer science etc., whose applications span diverse domains in both industrial and daily contexts [1]. As science and technology progress continuously, there is an escalating demand for enhancing stability and robustness in detecting dense small targets. Overcoming real-world challenges, such as mitigating interference from complex backgrounds and mitigating the impact of disturbances on detection, represents critical issues in object detection tasks [2]. In today's landscape, unmanned aerial vehicles (UAVs) are increasingly deployed in various global domains, elevating UAV-based aerial image recognition and detection to a prominent research frontier. Aerial images are utility in forestry and agricultural crop detection [3], [4], [5], intelligent city transportation [6], urban planning [7], municipal management [8], [10], power line inspection [10], [11], emergency rescue operations [12], [13], [14], among other applications. Nevertheless, the detection of small targets in aerial images remains a significant challenges due to some factors such as lighting variations, angles, and obstructions. Additionally, high exposure and complex backgrounds for detecting small targets in images substantially intensify the difficulty, thereby imposing greater demands on current object detection algorithms.

Aerial images, captured from an overhead perspective, inherently exhibit greater complexity and encompass a higher abundance of small objects compared to other image categories. The definition of small targets in object detection is approached using two methods including relative scale-based definition [15] and absolute scale-based definition [16]. In the relative scale-based approach, objects whose bounding box width and height are one-tenth of the original image width and height, respectively, are deemed small targets. The absolute scale-based approach designates targets with pixel values ranging from 20 to 32 as small targets, while targets with pixel values ranging from 2 to 20 are further classified as tiny targets. Given the intricacies of real-world aerial scenes, the proportion of small targets in the images is usually limited, resulting in scarce information available for small target detection and

DOI: 10.13164/re.2024.0012

thus exacerbating the detection difficulty. Presently, deep learning algorithms have yet to deliver a satisfactory solution to these challenges, underscoring the significance of theoretical research on the automatic detection of small-sized and densely-distributed objects in aerial images.

YOLOv5, a widely adopted model in the You Only Look Once (YOLO) series, offers a lightweight design suitable for efficient deployment in various environments. In this study, a detection-based multi-object detection algorithm built upon YOLOv5 is proposed. Our approach leverages the traditional YOLOv5 multi-object detection algorithm with the proposed DIOU-NMS, and extensive experiments are conducted on the VisDrone2019 dataset to demonstrate the effectiveness and feasibility of the proposed algorithm. The main contributions of this research are as follows.

(1) For addressing the issue of missing small-scale targets, an additional upsampling operation after the two existing upsampling operations in the YOLOv5s model structure is introduced, which results in a 160 × 160 feature map, then fusing with the second-layer feature map of the backbone network to obtain a larger feature map specifically suited for detecting small objects. This modification significantly enhances the detection accuracy compared to the original classic model.

(2) Furthermore, the Distance Intersection Over Union loss Non-Maximum Suppression (DIOU-NMS) method is proposed as a replacement for the conventional Non-Maximum Suppression (NMS) for candidate box filtering. DIOU-NMS effectively enhances the ability to capture small targets, leading to further improvements on detection accuracy.

The remainder of this paper is organized as follows. In Sec. 2 an overview of recent research on target detection algorithm is presented. In Sec. 3, detailed explanations of the specific implementation process of our proposed multi-object detection method are provided. Section 4 covers the dataset, evaluation metrics, and experimental setup details. The experimental results and comparative analysis are given in Sec. 5. Finally, in Sec. 6, our research findings and outline future directions for potential studies are concluded.

# 2. Related Research

Object detection aims to locate the positions of targets, obtain bounding boxes around them, and then extract appearance features to recognize their categories, thereby determining the identity of the targets. In simple terms, it involves accurately locating objects in a video frame, identifying their categories, and obtaining their position coordinates and size.

## 2.1 Research Status of Object Detection

The field of object detection has a history of several decades and can be roughly divided into two stages. The first stage consists of traditional object detection algorithms, which primarily relied on manually designed object features. These methods involved selecting potential regions of interest containing the objects, extracting features from these regions, and performing feature classification to achieve object detection. For example, Viola and Jones proposed the Viola-Jones detection algorithm [1], [17], which utilized Haar-like wavelet features and integral image techniques with AdaBoost for object detection using sliding windows. Another significant contribution was the histogram of oriented gradients (HOG) feature detection algorithm introduced by Dalal and others [18]. It represented the local appearance and shape of objects using the density distribution of gradients or edges, providing stability against geometric deformations and lighting variations, and laying an important foundation for subsequent detection methods. In 2008, Felzenszwalb et al. [19] proposed the deformable parts model (DPM), which improved upon the HOG features. DPM enhanced detection accuracy by hard negative mining, bounding box regression, and context modeling. However, it exhibited limited stability when dealing with significant object rotations.

As computer hardware capabilities continued to improve, the demand for both accuracy and real-time performance in object detection increased. Traditional object detection algorithms could not cope with the vast amounts of data present in images and videos. The advent of deep learning brought about a new opportunity for enhancing the performance of object detection algorithms, ushering in the second stage of development-research based on deep learning. Deep learning is a class of multi-layer neural network algorithms that automatically learn hidden information from training data, transforming pixel data from images or videos into higher-order, more abstract features. Compared to traditional object detection methods, deep learning-based algorithms exhibit remarkable speed, high accuracy, and strong robustness. Deep learning-based object detection algorithms can be categorized into two branches including two-stage object detection algorithms and one-stage object detection algorithms.

The two-stage object detection algorithm initially generates candidate regions from the input image, followed by the generation of target bounding boxes within these identified candidate regions. An exemplar of the two-stage approach is the region convolutional neural networks (R-CNN) introduced by Girshick et al. [20]. This methodology employs the selective search (SS) technique to delineate potential candidate boxes within the image, hypothesizing the presence of objects. Subsequently, these candidate boxes are uniformly resized to a predetermined dimension and input into a convolutional neural network (CNN) architecture to facilitate the extraction of discriminative features. The extracted feature representations are then supplied as input to a support vector machine (SVM) for the purpose of classifying and predicting the presence of target objects within the candidate boxes. Furthermore, the SVM aids in predicting the specific category to which the detected objects belong. Building upon R-CNN, He et al. [21] pro-

posed the spatial pyramid pooling net (SPPNet) algorithm, which incorporated the spatial pyramid pooling (SPP) layer. This innovation allowed the computation of fixed-size feature maps for the entire image, eliminating the network's dependence on input image size and avoiding redundant computation of convolutional feature maps. Girshick et al. [22] further advanced the field with the fast R-CNN algorithm, amalgamating the strengths of R-CNN and SPPNet, enabling end-to-end training and achieving improved detection performance. In addition, Zhang et al. [23] introduced the faster R-CNN algorithm, which addressed the time-consuming nature of the SS method by employing clustering and constructing a region proposal network (RPN) for region extraction, classification, and regression. This advancement paved the way for real-time object detection.

In contrast, one-stage object detection algorithms bypass the candidate region generation phase, and directly predict object class probabilities and bounding box coordinates in a single step, resulting in faster detection speed. YOLOv1 [24] emerged as a pioneering one-stage object detection algorithm, which partitioned the image into multiple grids, and simultaneously predicted class probabilities and bounding boxes for each grid, achieving higher detection speed while maintaining better accuracy compared to the two-stage R-CNN algorithm. Subsequently, Redmon et al. [25] introduced YOLOv3, which further enhanced the detection capability for small objects by incorporating multi-scale predictions and redesigning the loss function. Bochkovskiy et al. [26] proposed the YOLOv4 algorithm, optimizing the YOLOv3 structure and achieving remarkable levels of accuracy and speed. In 2020, Ultralytics [27] further advanced the YOLO series with the YOLOv5 algorithm, which introduced the focus operation and integrated the CSP2 structure from the lightweight network Cross Stage Partial Network (CSPNet) [28] into the bottleneck network, enhancing feature fusion and further improving detection performance. The YOLO series of algorithms has undergone continuous development and refinement, solidifying its status as one of the most widely used object detection methods. Consequently, in this paper an in-depth investigation of the YOLOv5 object detection algorithm is undertaken.

## 2.2 Small Object Detection Algorithm

In the field of object detection with deep learning, the detection of small objects has become a hot research topic due to their small proportion in images and lack of distinctive features. This challenge is particularly prominent in aerial image detection, where complex backgrounds, relatively low resolutions, a high prevalence of small objects, overlapping and occlusion, and a top-down perspective further complicate detection. As a result, research in this area has emerged as a crucial subfield on small object detection.

Early on, researchers like Lin et al. proposed the use of feature pyramids [29] to address scale variations in images by fusing features from different hierarchical levels and constructing feature pyramids. However, this approach had limitations in preserving feature information for small objects. In response, Deng et al. [30] introduced an extended feature pyramid network (EFPN) designed specifically for small object detection. Yang et al. [31] introduced QueryDet, a novel query mechanism, to accelerate the inference speed of feature pyramid-based object detectors.

In recent years, the YOLO series of detection methods have gained wide adoption in unmanned aerial vehicle (UAV) image-based object detection due to their speed and accuracy. Liu et al. [11] introduced MTI-YOLO for tasks like inspecting insulators on power lines using drones. Zhou et al. [32] presented YOLOv3 with squeeze excitation for small object detection in remote sensing images, reducing computational costs. Recognizing the limited representation of small objects after multiple downsampling steps and their potential submergence in the background, Min et al. [33] proposed the FE-YOLOv5 model with a feature enhancement module (FEM) and spatially aware module (SAM) to capture detailed semantic and foreground information.

Kim and fellow researchers [34] introduced ECAP-YOLO, an efficient channel attention pyramid YOLO model, for small object detection in aerial images, leveraging efficient pyramid channel attention to enhance the YOLO backbone. Luo et al. [35] improved YOLOv5 by incorporating a feature extraction module with three asymmetric convolutions to strengthen the extraction of less prominent features. Pei et al. [36] developed the LCB-YOLOv5 model, which consists of a new module composed of lightweight stable modules (LSM) and a cross-stage partial network with three convolution (C3) structure modules to extract multiple features of small objects.

The YOLO series of algorithms, continually evolving and improving, have now become one of the most widely applied methods for small object detection.

## 2.3 YOLOv5 Object Detection Algorithm

The most significant difference between the YOLO series object detection algorithms and previous object detection methods at the time lies in the elimination of the two-stage process, comprising region proposals and classification. YOLO focuses on a one-stage object detection approach. The original YOLOv1 was an adaptation of the GoogLeNet structure [36], composed of 24 convolutional layers and two fully connected layers. To address the problems of YOLOv1's lower detection rates and larger detection errors, YOLOv2 was introduced. In YOLOv2 anchor boxes were incorporated, and the training was adjusted by predicting target object types as many classes as possible. It employed DarkNet-19 as the backbone, featured a deeper network structure with $3 \times 3$ convolutional kernels, removed dropout from convolutional layers [37], used batch normalization [38], and transformed the final layer into a convolutional layer. Additional techniques such as skip connections and multi-scale predictions were implemented, resulting in improved accuracy and robustness.

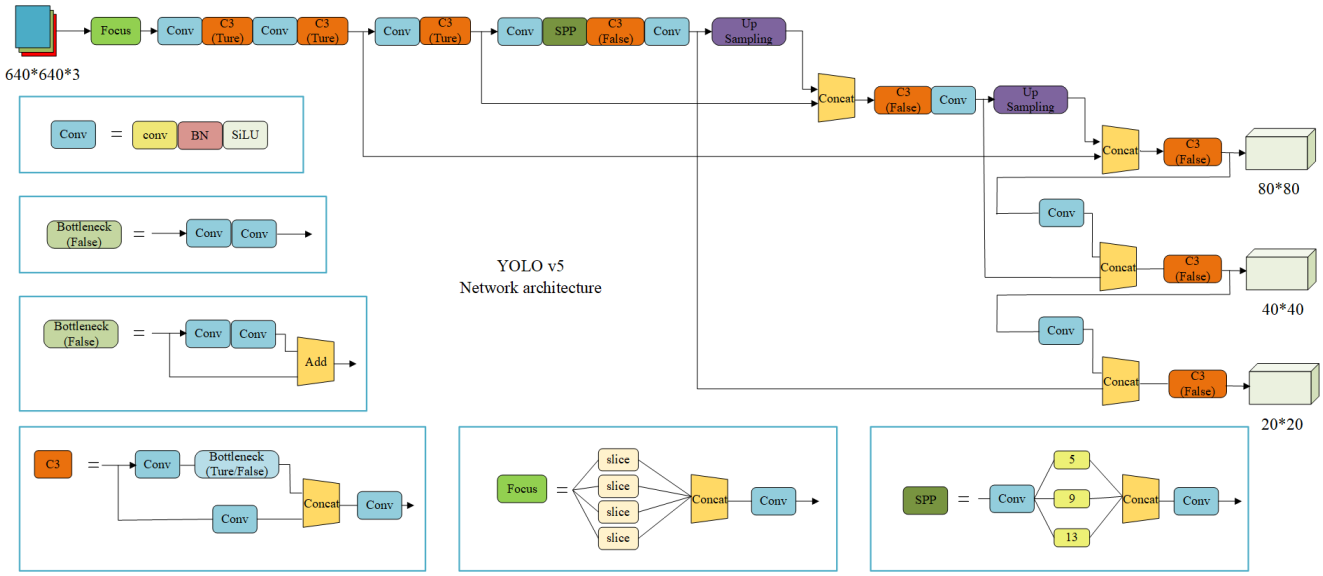


**Fig. 1.** The framework of YOLOv5 algorithm.

In YOLOv3, the network architecture was further refined with DarkNet-53 as the backbone, and it introduced pyramid-like feature extraction layers and integrated residual network structures, leading to substantial accuracy improvements. It adopted a multi-label classification and binary cross-entropy loss [39] for bounding boxes. To obtain semantic information from preceding layers and fine-grained details from the initial layers, the network concatenated feature maps extracted at the beginning with upsampled feature maps three times before making predictions at the final scale. Bounding box clustering techniques were employed to effectively address the problem of imbalanced object class numbers.

YOLOv4 utilized CSPDarknet-53 [28] as the main network, featuring the cross stage partial (CSP) structure, which sped up training and enhanced detection accuracy. It introduced the path aggregation network (PANet) [40] module for combining features from different levels, effectively improving small object detection capability and enhancing object localization accuracy. The inclusion of the spatial attention module (SAM) [41] enabled adaptive feature map importance adjustments.

YOLOv5, building upon the previous versions, incorporated the DropBlock regularization technique [42] to reduce redundant information in feature maps, improving the model's generalization. Model simplification enhanced detection speed. It also supported exporting pre-trained models in open neural network exchange (ONNX) format, enabling model deployment in various deep learning frameworks, making it more adaptable for aerial and UAV-based object detection tasks.

This research is based on the one-stage YOLOv5 object detection algorithm, which is designed to identify object locations and sizes in a given video frame, excluding background information. The overall architecture of the YOLOv5 algorithm is depicted in Fig. 1.

The YOLOv5 architecture is comprised of four key components including the input module, backbone network, neck network, and output module. In the input module, the original image is subjected to preprocessing, and data augmentation is performed using the mosaic method, which enables adaptive image scaling and anchor box calculation. The backbone network encompasses modules such as focus, convolutional block lightweight (CBL), and cross-stage partial (CSP) structures, which are convolutional neural networks designed to extract features of varying granularity from the image. The neck network combines the feature pyramid networks (FPN) and path aggregation networks (PAN) to enhance feature fusion. FPN facilitates the propagation of semantic information from top to bottom, while PAN facilitates the propagation of positional information from bottom to top, thus bolstering the network's capacity for effective feature fusion. In the output module, the processed image features are subject to predictions at three different scales to generate bounding boxes and predict the class of the target object. Subsequently, the non-maximum suppression (NMS) method is applied to filter and retain the most pertinent bounding boxes.

## 3. Research Methods

The YOLOv5 object detection algorithm is divided into four models, namely YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, based on variations in convolutional depth and the number of convolutional kernels. These models offer increasing detection accuracy as their size grows. However, their real-time performance declines due to increased complexity. To strike a balance between accuracy and real-time requirements, in the study the YOLOv5s model is adopted as the foundation for further improvements, and an enhanced multi-object detection algorithm based on DIOU-NMS is proposed.

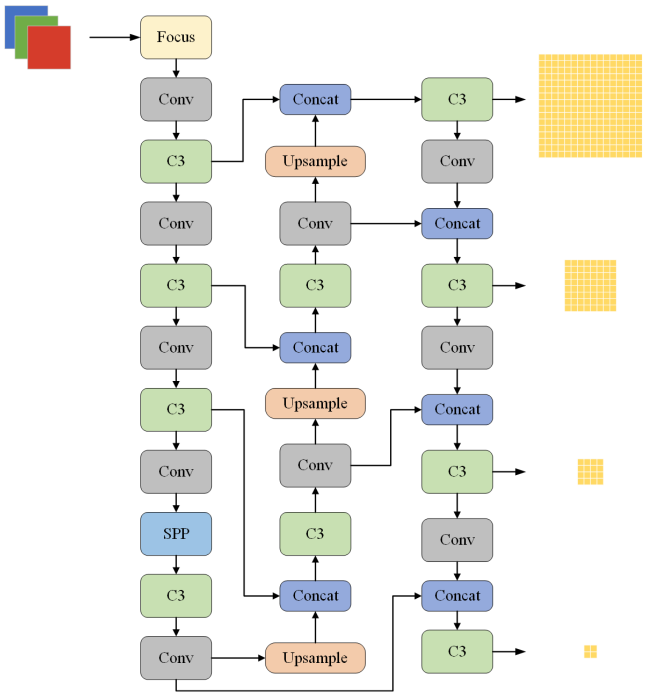


**Fig. 2.** Improved structure of YOLOv5.

In the proposed algorithm two key enhancements are introduced. Firstly, it incorporates a small object detection layer, which involves an additional upsampling operation to improve the detection capability for small objects. Secondly, the traditional non-maximum suppression (NMS) method used in YOLOv5 is replaced with DIOU-based NMS, i.e., DIOU-NMS. This modification addresses the issue of missed detections that may arise in dense object scenarios. The improved network structure is shown in Fig. 2.

## 3.1 Small Target Detection Layer

The original YOLOv5 algorithm takes input images with a size of 640 × 640 pixels, and employs three detection layers at 8, 16, and 32 times downsampling locations to predict features at three scales. Consequently, the feature map sizes after the three downsampling stages are 80 × 80, 40 × 40, and 20 × 20 pixels, respectively, corresponding to the detection of objects in 8 × 8, 16 × 16, and 32 × 32 regions. However, it exhibits limitations for effectively detecting small, densely packed objects. In aerial image object detection tasks, numerous distant objects exist, occupying only a few pixels in the entire video frame and containing minimal information, thereby making it challenging to extract meaningful features for accurate detection.

To address this limitation, an additional detection layer is incorporated into the YOLOv5 architecture to enhance its tolerance for small-scale objects. Specifically, after two rounds of upsampling in the original network structure, an extra upsampling operation is performed, resulting in a feature map size of 160 × 160 pixels which is then fused with the second layer feature map to obtain a larger receptive field, thus enabling improved detection of small objects. In the end, the network utilizes four detection layers to achieve multi-scale object detection.

## 3.2 DIOU-NMS

In conventional NMS algorithms, the detection box with the highest confidence score is compared with other detected boxes individually to compute their intersection over union (IOU). The IOU is obtained by dividing the area of intersection between the predicted box and the ground truth box by the area of their union. If the IOU value exceeds a predefined threshold, the corresponding detection box is deemed redundant, and filtered out. It is apparent that, in traditional NMS, IOU is the sole consideration and serves as the exclusive criterion for evaluating detection success.

However, IOU alone has its limitations, as it solely examines the relationship between the intersecting and union areas of two detection boxes. This may result in cases where it fails to serve as an effective filtering criterion. For instance, as illustrated in Fig. 3, with red representing the ground truth box, and blue and green denoting detection boxes, Figure 3(a) shows that when both detection boxes are spatially distant from the ground truth box, the IOU value is 0, rendering them unfiltered.

Figure 3(b) means that two detection boxes have identical sizes and intersect the ground truth box by the same area, the IOU values are equal, and they evade filtering. Figure 3(c) denotes that similarly, when two detection boxes share the same size, and are entirely encompassed within the ground truth box, the IOU values are identical, precluding effective filtering.

In response to the aforementioned limitations of NMS, a novel DIOU-NMS is proposed, which introduces the distance intersection over union (DIOU) calculation method. DIOU-NMS replaces the conventional IOU computation in the traditional NMS with DIOU. As depicted in Fig. 4, DIOU takes into account both the overlapping area and the distance between the centers of two detection boxes, enabling more accurate selection of detection boxes and facilitating faster convergence. The expression for DIOU is given by (1).

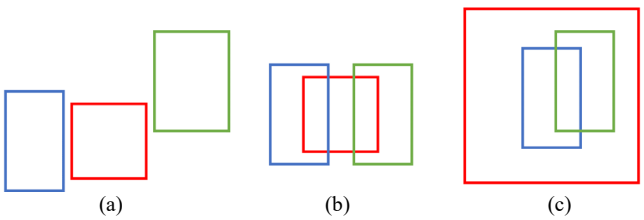


**Fig. 3.** Cases that cannot be recognized by NMS. Red represents the ground truth bounding box, while blue and green represent the detected bounding boxes. (a) The detection frame is separated from the real frame; (b) the size of the detection frame is the same and the area that intersects the real target frame is also the same; (c) the detection frame is the same size and both are located inside the real target frame.

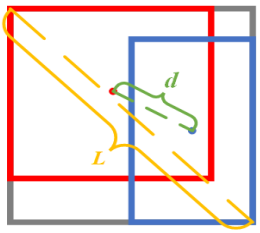


**Fig. 4.** DIOU schematic where red represents the ground truth bounding box, while blue represents the detection box.

$$DIOU = 1 - IOU + \frac{d^2(A,B)}{L^2} \tag{1}$$

where $d$ represents the distance between the ground truth box A and the detection box B. $L$ denotes the diagonal length of the minimum enclosing box that encompasses both the detection box and the ground truth box. The term "*IOU*" refers to the intersection over union, which is computed by dividing the area of intersection between the predicted box and the ground truth box by the area of their union.

# 4. Experiment Setup

## 4.1 Experiment Configuration

The hardware configuration for the experiments in this paper includes an Intel Core i7-10875H 2.30 GHz CPU, an NVIDIA GeForce RTX2060 GPU, and 16GB RAM. The environment setup comprises python 3.8, CUDA 11.3, and Pytorch 1.11.0.

## 4.2 Measure Datasets

The experiments in this study were conducted using the VisDrone2019 [43] small object dataset, collected by the AISKYEYE team in the Machine Learning and Data Mining Laboratory of Tianjin University. The dataset comprises 288 video clips captured by unmanned aerial vehicles (UAVs) under various scenes such as weather conditions, and lighting conditions, consisting of over 260,000 video frames and 10,000 static images. The dataset is characterized by an abundance of small and occluded objects, posing significant challenges to object detection algorithms, leading to potential issues of false positives and missed detections. Some examples of the VisDrone2019 dataset are given in Fig. 5.

Given the dataset's high image resolution, with image sizes of $6000 \times 4000$ pixels, image resizing becomes a necessary preprocessing step. However, resizing may further diminish already small objects to almost negligible sizes, making it challenging for the algorithm to learn meaningful features, significantly impacting the training effectiveness. To address this problem, a segmentation approach is employed, dividing each dataset image into six smaller blocks using a $2 \times 3$ grid pattern. Nonetheless, during the segmentation process, some objects may precisely lie on

**Fig. 5.** Some examples of the VisDrone2019 dataset.

the edges between two sub-images, potentially causing object truncation and introducing artificial interference in the detection results. To mitigate such issues, a 20% overlapping area is introduced between the segmented sub-images, and the bounding box coordinates are adjusted accordingly, thus alleviating the impact on detection performance during the dataset image segmentation process.

# 5. Experimental Results and Analysis

## 5.1 Evaluation Indicators

The effectiveness and feasibility of the algorithm improvements are validated using four evaluation metrics including Precision, Recall, Mean Average Precision (mAP) at IOU confidence threshold of 0.5 (mAP@0.5), and mAP at IOU confidence threshold ranging from 0.5 to 0.95 (mAP@0.5:0.95).

*Precision*, also known as positive predictive value, represents the ratio of correctly detected samples to the total number of samples detected. It is computed as shown in (2):

$$Precision = \frac{TP}{TP + FP}. \tag{2}$$

*Recall*, also known as sensitivity or true positive rate, represents the ratio of correctly detected samples to the total number of samples that should have been detected. It is computed as shown in (3):

$$Recall = \frac{TP}{TP + FN}. \tag{3}$$

In this context, true positive ($TP$) represents the positive samples that the model correctly predicts as positive,

| The real situations | The predicted situations | |
|---|---|---|
| | Positive sample | Negative sample |
| Positive sample | TP | FN |
| Negative sample | FP | TN |

**Tab. 1.** Detailed description of the parameters of evaluation indicators.

while true negative (*TN*) represents the negative samples that the model correctly predicts as negative. On the other hand, false positive (*FP*) denotes the negative samples that the model incorrectly predicts as positive, and false negative (*FN*) represents the positive samples that the model incorrectly predicts as negative. The details are illustrated in Tab. 1.

Average precision (*AP*) is the area under the Precision-Recall (P-R) curve, which is obtained by plotting *Precision* against *Recall*. It quantifies the overall performance of an object detection model. The mean average precision (*mAP*) is the average value of *AP* across all object detection categories.

## 5.2 Ablation Experiments

For verifying the effectiveness of the proposed ideas on small target detection layer and DIOU-NMS, the results of the ablation experiments are presented in Tab. 2.

From the results of the conducted ablative experiments, as presented in Tab. 2, it becomes apparent that the incorporation of a small object detection layer yields noteworthy enhancements in the network model's performance. The model achieves this by adeptly fusing deep-level features with shallow-level features via an upsampling mechanism. This strategic integration empowers the network to discern minute object features within images with heightened precision. Consequently, an elevation of 6 percentage points in accuracy, 5.2 percentage points in recall, 9.4 percentage points in mAP@0.5, and 7.3 percentage points in mAP@0.5:0.95 is observed. This collective improvement underscores the substantial mitigation of false positives and false negatives for diminutive objects following the integration of the small object detection layer.

Subsequently, an insightful juxtaposition is drawn with the incorporation of the DIOU-NMS algorithm. The conventional Non-Maximum Suppression (NMS) technique predominantly relies on the Intersection over Union (IOU) paradigm to adjudicate detection boxes. This process entails determining the most prominent detection box predicated upon the relative magnitude between the intersected area and the union area of two detection boxes. However, this approach often demonstrates inadequacy in effectively winnowing out undesired outcomes under diverse circumstances. Thus, the introduction of DIOU-NMS emerges as a salient remedy. By encapsulating both the overlap area and the center point distance between two detection boxes, this innovative method optimizes box selection, culminating in outcomes of higher precision and expedited convergence. In tandem with the outcomes presented in Tab. 2, an incremental augmentation of 0.2 percentage points in accuracy, 1.5 percentage points in recall, 1.7 percentage points in mAP@0.5, and 3.2 percentage points in mAP@0.5:0.95 is manifest. Although these enhancements may not be deemed monumental in isolation, their culmination with the small object detection layer precipitates a cumulative escalation of 7.8 percentage points in accuracy, 9.8 percentage points in recall, 10.9 percentage points in mAP@0.5, and 10.1 percentage points in mAP@0.5:0.95. This palpably underscores the symbiotic harmony between DIOU-NMS and the small object detection layer. The concerted interplay of these improvements exhibits an emergent synergy, whereby the composite effect exceeds the mere summation of individual parts, thereby rendering a constructive contribution to the efficacy of object detection.

Finally, a comprehensive comparative assessment is conducted, juxtaposing the proposed algorithm with the antecedent enhancements of the YOLO algorithm. Table 2 efficaciously portrays a discernible amelioration in accuracy. Relatively, when contrasted with the incremental impact of integrating the small object detection layer, an amelioration of 1.8 percentage points is observed. However, when juxtaposed with the assimilation of DIOU-NMS, an appreciable enhancement of 7.6 percentage points is substantiated. Notably, an appreciable elevation in mAP is also evident. Cross-referencing Figure 6 underscores the demonstrably better detection acumen of the algorithm posited within this exposition, as compared to the exclusive integration of DIOU-NMS. This superiority particularly resonates with the identification of diminutive objects such as pedestrians and bicycles within images. Likewise, a comparative analysis vis-à-vis the exclusive integration of the small object detection layer reveals the algorithm's heightened efficacy in detecting objects situated at a considerable distance or partially obscured by intervening obstacles.

| Models | Precision | Recall | mAP@0.5 | mAP@0.5: 0.95 |
|---|---|---|---|---|
| YOLOv5 | 0.817 | 0.717 | 0.801 | 0.452 |
| YOLOv5+ Small target detection layer | 0.877 | 0.769 | 0.895 | 0.535 |
| YOLOv5+DIOU-NMS | 0.819 | 0.732 | 0.818 | 0.484 |
| The proposed algorithm | 0.895 | 0.815 | 0.910 | 0.553 |

**Tab. 2.** Ablation experimental results of VisDrone2019 dataset.


**Fig. 6.** Comparison of ablation experimental results.

| Algorithm | Precision | Recall | mAP@0.5 | mAP@0.5: 0.95 |
|---|---|---|---|---|
| YOLOv5 | 0.817 | 0.717 | 0.801 | 0.452 |
| Fast R-CNN | 0.874 | 0.796 | 0.886 | 0.539 |
| Faster R-CNN | 0.863 | 0.813 | 0.893 | 0.555 |
| Proposed in this paper | 0.895 | 0.815 | 0.910 | 0.553 |

**Tab. 3.** Comparison results of tracking performance on VOT2018 dataset.

## 5.3 Comparative Experiments

The comparative experiments were conducted by selecting the fundamental algorithm YOLOv5 and two-stage object detection algorithms renowned including fast R-CNN and faster R-CNN. The results are presented in Tab. 3.

From the quantitative comparative results presented in Tab. 3, it is evident that the enhanced YOLOv5 algorithm falls slightly short of the two-stage object detection algorithm, Faster R-CNN, only in the mAP@0.5:0.95 metric by a marginal margin of 0.2 percentage points. However, in all other metrics, it demonstrates superiority, underscoring the more advanced detection performance of the algorithm introduced in this study. Faster R-CNN is a region-based convolutional neural network (CNN) object detection methodology, employing a two-stage detection strategy. In its initial phase, it leverages a Region Proposal Network (RPN) to generate prospective object boxes, followed by object classification and precise localization within these candidates. This bifurcated approach excels in accuracy due to its improved object localization precision. Correspondingly, as discerned from Tab. 3, Faster R-CNN closely approximates the proposed algorithm across metrics such as Recall and mAP.

However, it is pertinent to note that this approach's computational efficiency is relatively compromised, particularly when applied to extensive datasets. By contrast, the approach posited in this research embodies a singular-stage object detection algorithm, inspired by the "You Only Look Once" (YOLO) principle. Herein, object detection is conceived as a regression conundrum, whereby predictions pertaining to object bounding boxes and class attributions are concurrently computed to achieve detection. Significantly, the incorporation of DIOU-NMS introduces considerations of distance and overlapping area amongst object boxes during the Non-Maximum Suppression (NMS) phase. This evolved NMS mechanism facilitates the culling of meaningful detection boxes, thereby concretizing ad-

vancements in both precision and recall. The algorithm, attuned to the characteristics of diminutive entities, further integrates a small object detection stratum. This augmentation holds pivotal import for the precision of small object detection within aerial imagery, wherein objects often tend to be in close proximity, subject to deformation, or even overlapping. The confluence of the small object detection stratum with DIOU-NMS is instrumental in augmenting the model's prowess in such scenarios. This effect is reinforced through the assimilation of an additional upsampling module during the feature map integration process, thereby ameliorating the model's capacity to encapsulate shallow-level features. Consequently, a remarkable prowess in small object detection is witnessed, culminating in noteworthy enhancements across metrics, most notably mAP@0.5. Visual corroboration in Fig. 7 substantiates this contention, unmasking more advanced detection of petite objects, such as bicycles and pedestrians, when juxtaposed with Faster R-CNN.

The ascendancy of YOLO emanates from its singular-stage detection strategy, which simplifies object detection into a regression problem. Through a solitary pass, both object bounding box predictions and class ascriptions are simultaneously surmised, thereby conferring a pivotal computational speed edge. This vantage point is particularly pronounced when the algorithm is deployed in datasets featuring dense object distributions, exemplified by Visdrone.

The preeminence of the proposed algorithm vis-à-vis the Visdrone aerial dataset harboring compact small objects is underpinned by a trifecta of factors: its frugal singular-stage design, assimilation of DIOU-NMS, and optimization catering to the exigencies of small object detection. Augmenting these attributes is the algorithm's end-to-end training approach, synergistically amalgamating to yield commendable performance across critical metrics such as Precision, Recall, and mAP@0.5. While Faster R-CNN preserves fidelity across select metrics, its overarching performance is somewhat curtailed by the constraints inherent to its two-stage design. A visual juxtaposition in Fig. 7 furnishes a tangible insight into the comparative performance of the proposed algorithm vis-à-vis Faster R-CNN on the VisDrone2019 dataset, vividly elucidating the algorithm's propensity for more effective detection of diminutive and densely distributed objects.

# 6. Conclusion and Future Work

In this paper an introduction to the YOLOv5 object detection algorithm is presented, and an enhanced version, termed the YOLOv5 multi-object detection approach, using DIOU-NMS is proposed. The motivation behind this improvement is to effectively address scenarios involving small-scale and densely packed targets, thereby better serving subsequent multi-object tracking tasks. The primary enhancements involve extending the YOLOv5s model with an additional detection layer specifically designed for small objects for achieving an excellent performance by the introduction of an upsampling operation. This modification empowers the algorithm to perform exceptionally well in small object detection tasks. Furthermore, the traditional NMS method is replaced by the proposed DIOU-NMS, which significantly enhances the selection of detection boxes, resulting in more precise outcomes, and mitigating issues related to target omission caused by excessively dense target regions.

Experimental results on the VisDrone2019 dataset demonstrate that the proposed YOLOv5 multi-object detection algorithm using DIOU-NMS effectively handles scenarios with small-scale and densely packed targets, yielding better detection performance and providing highly accurate detection outcomes. In the future, building upon the advancements made in multi-object detection using this study, we aim to concentrate on dynamic multi-object tracking research, seeking to further enhance the accuracy, robustness, and real-time capabilities of motion object detection and tracking algorithms, especially in complex backgrounds. Our objective is to continually optimize and refine the algorithm model, improve region discrimination performance, and foster the development of state-of-the-art multi-object tracking algorithms.

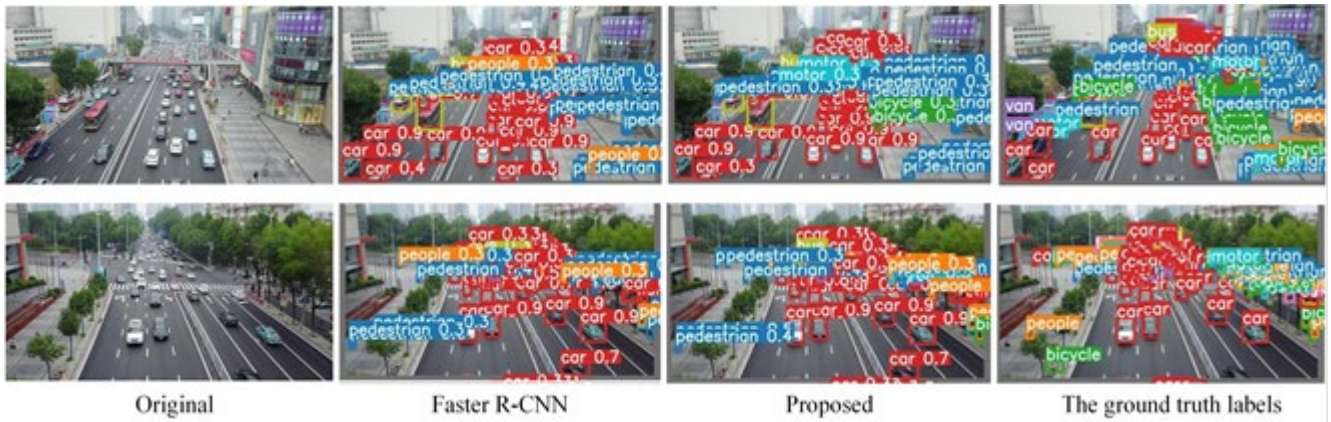


**Fig. 7.** Comparison between the algorithm proposed in this paper and the detection results of Faster R-CNN on moving small targets.

# Acknowledgments

This work is supported by the Joint Project of Beijing Natural Science Foundation and Beijing Municipal Education Commission (Grant No. KZ202110011015).


# References


[1] BUGEAU A., PÉREZ P. Detection and segmentation of moving objects in complex scenes. *Computer Vision and Image Understanding*, 2009, vol. 113, no. 4, p. 459–476. DOI: 10.1016/j.cviu.2008.11.005

[2] YIN, H., CHEN, B., CHAI, Y., et al. Review of vision-based object detection and tracking (in Chinese). *Acta Automatica Sinica*, 2016, vol. 42, no. 10, p. 1466–1489. DOI: 10.16383/j.aas.2016.c150823

[3] OSCO, L. P., DE ARRUDA, M. D. S., GONCALVES, D. N., et al. A CNN approach to simultaneously count plants and detect plantation-rows from UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing,* 2021, vol. 174, p. 1–17. DOI: 10.1016/j.isprsjprs.2021.01.024

[4] SIVAKUMAR, A. N. V., LI, J. T., SCOTT, S., et al. Comparison of object detection and patch-based classification deep learning models on mid- to late-season weed detection in UAV imagery. *Remote Sensing*, 2020, vol. 12, no. 13, p. 2136–2140. DOI: 10.3390/rs12132136

[5] WANG, L., XIANG, L. R., TANG, L., et al. A convolutional neural network-based method for corn stand counting in the field. *Sensors*, 2021, vol. 21, no. 2, p. 507–510. DOI: 10.3390/s21020507

[6] AMMOUR, N., ALHICHRI, H., BAZI, Y., et al. Deep learning approach for car detection in UAV imagery. *Remote Sensing*, 2017, vol. 9, no. 4, p. 312–316. DOI: 10.3390/rs9040312

[7] LIU, Y., SHI, G., LI, Y., et al. M-YOLO based detection and recognition of highway surface oil filling with unmanned aerial vehicle. In *Proceeding of 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*. Xi'an (China), 2022, p. 1884–1887. DOI: 10.1109/ICSP54964.2022.9778782

[8] DING, W., ZHANG, L. Building detection in remote sensing image based on improved YOLOV5. In *Proceeding of 2021 17th International Conference on Computational Intelligence and Security (CIS).* Chengdu (China), 2021, p. 133–136. DOI: 10.1109/CIS54983.2021.00036

[9] ZHANG, R., WEN, C. SOD-YOLO: A small target defect detection algorithm for wind turbine blades based on improved YOLOv5. *Advanced Theory and Simulations*, 2022, vol. 5, no. 7, p. 2100631–2100635. DOI: 10.1002/adts.202100631

[10] GUO, J., XIE, J., YUAN, J., et al. Fault identification of transmission line shockproof hammer based on improved YOLO V4. In *Proceeding of International Conference on Automation and Applications (ICAA).* Nanjing (China), 2021, p. 826–833. DOI: 10.1109/ICAA53760.2021.00151

[11] LIU, C. Y., WU, Y. Q., LIU, J. J., et al. MTI-YOLO: A light-weight and real-time deep neural network for insulator detection in complex aerial images. *Energies*, 2021, vol. 14, no. 5, p. 1–19. DOI: 10.3390/en14051426

[12] SAMBOLEK, S., IVASIC-KOS, M. Automatic person detection in search and rescue operations using deep CNN detectors. *IEEE Access,* 2021, no. 9, p. 37905–37922. DOI: 10.1109/access.2021.3063681

[13] BOZIC-STUTIC, D., MARUSIC, Z., GOTOVAC, S. Deep learning approach in aerial imagery for supporting land search and rescue mission. *International Journal of Computer Vision*, 2019, vol. 127, no. 9, p. 1256–1278. DOI: 10.1007/s11263-019-01177-1

[14] DE OLIVEIRA, D. C., WEHRMEISTER, M. A. Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV. *Sensors*, 2018, vol. 18, no. 7, p. 1–33. DOI: 10.3390/s18072244

[15] CHEN, C., LIU, M. Y., TUZEL, O., et al. R-CNN for small object detection. In *Proceeding of Asian Conference on Computer Vision (ACCV)*. Taipei (Taiwan), 2016, vol. 5, p. 214–230. DOI: 10.1007/978-3-319-54193-8_14

[16] YU, X., GONG, Y., JIANG, N., et al. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Snowmass Village (USA), 2020, p. 1246–1254. DOI: 10.1109/WACV45572.2020.9093394

[17] VIOLA, P., JONES, M. Robust real-time face detection. *International Journal of Computer Vision*, 2004, vol. 57, no. 2, p. 137–154. DOI: 10.1023/B:VISI.0000013087.49260.fb

[18] DALAL, N., TRIGGS, B. Histograms of oriented gradients for human detection. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. San Diego (USA), 2005, p. 1–8. DOI: 10.1109/cvpr.2005.177

[19] FELZENSZWALB, P. F., GIRSHICK, R. B., MCALLESTER, D., et al. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, vol. 32, no. 9, p. 1627–1645. DOI: 10.1109/TPAMI.2009.167

[20] GIRSHICK, R., DONAHUE, J., DARRELL, T., et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus (USA), 2014, p. 580–587. DOI: 10.1109/CVPR.2014.81

[21] HE, K. M., ZHANG, X. Y., REN, S. Q., et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. In Fleet, D., Pajdla, T., Schiele, B., et al. (Eds.) *Computer Vision – ECCV 2014*, p. 346–361. DOI: 10.1007/978-3-319-10578-9_23

[22] GIRSHICK, R. Fast R-CNN. In *Proceedings of IEEE International Conference on Computer Vision.* Santiago (Chile), 2015, p. 1440–1448. DOI: 10.1109/ICCV.2015.169

[23] REN, S., HE, K., GIRSHICK, R., et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE *Transactions on Pattern Analysis and Machine Intelligence,* 2017, vol. 39, no. 6, p. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031

[24] REDMON, J., DIVVALA, S., GIRSHICK, R., et al. You Only Look Once: Unified, real-time object detection. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle (USA), 2016, p. 779–788. DOI: 10.1109/CVPR.2016.91

[25] REDMON, J., FARHADI, A. YOLOv3: An incremental improvement. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City (USA), 2018, p. 1–6. DOI: 10.48550/arXiv.1804.02767

[26] BOCHKOVSKIY, A., WANG, C. Y., LIAO, H. Y. M. YOLOv4: Optimal speed and accuracy of object detection. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle (USA), 2020. DOI: 10.48550/arXiv.2004.10934

[27] NELSON, J., SOLAWETZ J. *YOLOv5 is Here: State-of-the-Art Object Detection at 140 fps*. [Online] Available at: https://blog roboflow com/yolov5-is-here

[28] WANG, C. Y., LIAO, H. Y. M., WU, Y. H., et al. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Washington (USA), 2020, p. 1571–1580. DOI: 10.1109/CVPRW50498.2020.00203

[29] LIN, T. Y., DOLLAR, P., GIRSHICK, R., et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu (USA), 2017, p. 936–944. DOI: 10.1109/CVPR.2017.106

[30] DENG, C., WANG, M., LIU, L., et al. Extended feature pyramid network for small object detection. *IEEE Transactions on Multimedia*, 2021, vol. 24, p. 1968–1979. DOI: 10.1109/TMM.2021.3074273

[31] YANG, C., HUANG, Z., WANG, N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans (USA), 2022, p. 13658-13667. DOI: 10.1109/CVPR52688.2022.01330

[32] ZHOU, L., DENG, G., LI, W., et al. A lightweight SE-YOLOv3 network for multi-scale object detection in remote sensing imagery. *International Journal of Pattern Recognition and Artificial Intelligence*, 2021, vol. 35, no. 13. DOI: 10.1142/S0218001421500373

[33] WANG, M., YANG, W., WANG, L., et al. FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection. *Journal of Visual Communication and Image Representation*, 2023, vol. 90, p. 1–8. DOI: 10.1016/j.jvcir.2023.103752

[34] KIM, M., JEONG, J., KIM, S. ECAP-YOLO: Efficient channel attention pyramid YOLO for small object detection in aerial image. *Remote Sensing*, 2021, vol. 13, p. 1–20. DOI: 10.3390/rs13234851

[35] LUO, X., WU, Y., WANG, F. Target detection method of UAV aerial imagery based on improved YOLOv5. *Remote Sensing*, 2022, vol. 14, no. 19, p. 1–25. DOI: 10.3390/rs14195063

[36] SZEGEDY, C., LIU, W., JIA, Y., et al. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston (USA), 2015, p. 1–9. DOI: 10.1109/CVPR.2015.7298594

[37] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., et al. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 2014, vol. 15, no. 1, p. 1929–1958. DOI: 10.5555/2627435.2670313

[38] IOFFE, S., SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*. Lille (France), 2015, p. 448–456. DOI: 10.48550/arXiv.1502.03167

[39] ZHANG, Z., SABUNCU, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Montreal (Canada), 2018, p. 8792 to 8802. DOI: 10.48550/arXiv.1805.07836

[40] LIU, S., QI, L., QIN, H., et al. Path aggregation network for instance segmentation. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City (USA), 2018, p. 8759–8768. DOI: 10.48550/arXiv.1803.01534

[41] WOO, S., PARK, J., LEE, J. Y., et al. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Munich (Germany), 2018. DOI: 10.48550/arXiv.1807.06521

[42] GHIASI, G., LIN T. Y., LE, Q. V. DropBlock: A regularization method for convolutional networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Montreal (Canada), 2018, p. 10750–10760. DOI: 10.48550/ arXiv.1810.12890

[43] ZHU, P., WEN, L., DU, D., et al. Vision meets drones: Past, present and future. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle (USA), 2020. DOI: 10.48550/arXiv.2001.06303

# About the Authors ...

**Yu WANG** was born in 1977. She received her Ph.D. degree from the University of Science and Technology Beijing in 2009. She was engaged in scientific research as a post-doctoral in the Beijing Key Laboratory of Multidimensional and Multiscale Computing Photography, Tsinghua University from 2009 to 2011. She is now a Professor and doctoral supervisor of the Beijing Technology and Business University. Her research interests include pattern recognition, image processing, and computer vision.

**Xiang ZOU** was born in 1997. He is now a candidate of master degree in the School of Computer and Artificial Intelligence, Beijing Technology and Business University, China. His research interests include pattern recognition, image processing, and computer vision.

**Jiantong SHI** was born in 1996. He received his B.S. degree from Qingdao University of Science and Technology in 2019. He is now a candidate of master degree in the School of Artificial Intelligence, Beijing Technology and Business University, China. His research interests include pattern recognition, image processing, and computer vision.

**Minhua LIU** (corresponding author) was born in 1976. He received his Ph.D. degree from Tsinghua University. Now he is the vice president of Beijing Technology and Business University. His research interests include image processing and the modeling, and analysis of complex system.