International Journal of Computational Intelligence Systems (2022) 15:2
https://doi.org/10.1007/s44196-021-00056-3

RESEARCH ARTICLE

GLE-Net: A Global and Local Ensemble Network for Aerial Object Detection

Jiajia Liao[1] · Yujun Liu[1] · Yingchao Piao[2] · Jinhe Su[1] · Guorong Cai[1] · Yundong Wu[1]

Received: 29 April 2021 / Accepted: 10 December 2021

Abstract
Recent advances in camera-equipped drone applications increased the demand for visual object detection algorithms with deep learning for aerial images. There are several limitations in accuracy for a single deep learning model. Inspired by ensemble learning can significantly improve the generalization ability of the model in the machine learning field, we introduce a novel integration strategy to combine the inference results of two different methods without non-maximum suppression. In this paper, a global and local ensemble network (GLE-Net) was proposed to increase the quality of predictions by considering the global weights for different models and adjusting the local weights for bounding boxes. Specifically, the global module assigns different weights to models. In the local module, we group the bounding boxes that corresponding to the same object as a cluster. Each cluster generates a final predict box and assigns the highest score in the cluster as the score of the final predict box. Experiments on benchmarks VisDrone2019 show promising performance of GLE-Net compared with the baseline network.

Keywords Convolutional neural networks (CNNs) · Aerial images object detection · VisDrone2019 dataset · Deep learning · Ensemble algorithm

1 Introduction

Object detection in aerial images has become a challenging and active field in computer vision. Importantly, aerial object detection has been a significant success in many applications, i.e., disaster assistance, military, and agriculture. With the advancement of aerial photography techniques and equipment (e.g., unmanned aerial vehicles and satellites) to shoot high-resolution aerial images, more researchers have devised many object detection algorithms based on deep learning. The natural images often capture smaller visual fields and larger object sizes, whereas the aerial images generally capture the information of the lower resolution and small scale of the objects. Aerial images have a wide covered area and contain a mickle tiny and dense distribution of objects. Although many object detectors have achieved advanced performance on natural images, they are not able to attain satisfactory detection results on aerial images.

There are three special challenges for aerial images as follows: (1) aerial datasets mostly high-resolution images; (2) objects typically have small scales relative to the images, and (3) the object distribution of images is not uniform in large scenes. Therefore, it is difficult for the general-purpose detector to effectively detect objects of the aerial images and the most recent works focus on aerial images (e.g., CAD-Net [1] R^2CNN [2]), which cannot reach the level where the state-of-the-art object detection methods perform on natural images. To solve these

✉ Jinhe Su
sujh@jmu.edu.cn

Jiajia Liao
jiajialiao_08@163.com

Yujun Liu
yujunliu@jmu.edu.cn

Yingchao Piao
pyc@cnic.cn

Guorong Cai
guorongcai.jmu@jmu.edu.cn

Yundong Wu
yundongwu@jmu.edu.cn

[1] Computer Engineering College, Jimei University, 185 Yinjiang Rd., Jimei District, Xiamen 361021, China

[2] Computer Network Information Center, Chinese Academy of Sciences, Building No. 2, 4, Zhongguancun Nansijie, Haidian District, Beijing 100190, China

Published online: 03 January 2022

issues, a universal solution is to ensemble multiple weak detectors to form a robust and useful detector. The integrated machine learning model is a common method to improve models' capability, which has been used in many scenarios since it combines the decision of multiple models to upgrade the overall performance. These approaches have been effectively employed for improving accuracy in some machine learning tasks, and object detection is not an exception. Unfortunately, when it comes to the object detection model based on a deep neural network, it is not a simple process of merging detection results. Ensemble Methods [3] introduces several voting strategies to carry out the integration process and boosts the accuracy of many model object detection results. Inspired by the above strategy, we tested it with Yolov5 [4] and CenterNet [5, 6], respectively, as shown in Fig. 1. Our proposed method achieves the case of a lower missing rate and obtains higher accuracy performance.

In this paper, we propose a global and local ensemble network to enhance multi-model detection results for aerial image object detection, which can serve as an efficient plug-and-play network in existing scene parsing networks. Specifically, in our method, we select CenterNet [5, 6] and Yolov5 [4] as our basic model. Firstly, we trained Yolov5 and CenterNet on VisDrone2019 [7] dataset, separately. Secondly, using the prediction results of these two different models and inputting them into our proposed method (GLE-Net) to enhance detection performance by the global ensemble module and the local ensemble module.

To sum up, our work makes the following contributions:

- We propose a global and local ensemble network (GLE-Net) to integrate the inference results of multiple state-of-the-art detectors for object detection in aerial images.
- We design an effective plug-and-play module to fuse these predicted classification and box regression information of several detectors.
- Our method achieves better performance than the baseline pipeline models on aerial image dataset VisDrone2019 [7].

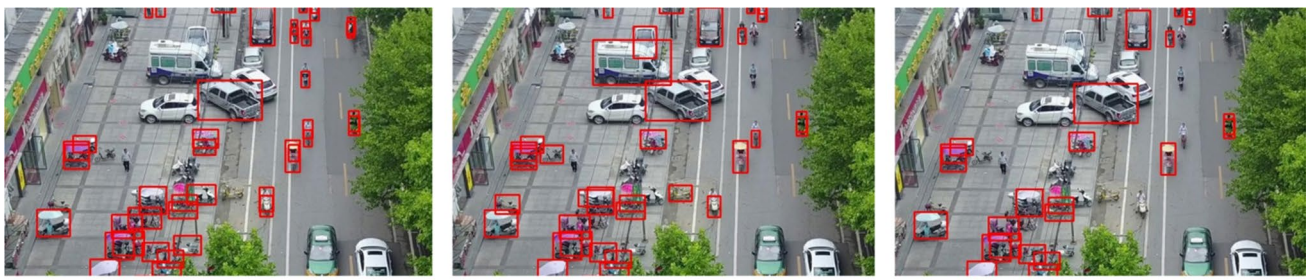
The rest of the paper is organized as follows: Sect. 2 introduces the related work about generic and aerial image object detection algorithms. Section 3 describes our proposed method in detail. Section 4 introduces datasets and experimental results. Section 5 is a summary of the paper.

2 Related Work

Object detection has received an important amount of attention in the last two decades. In this section, the most relevant work to ours is summarized under two subcategories: (1) Generic Object Detection and (2) Aerial Image Object Detection.

2.1 Generic Object Detection

With the rapid development of the deep neural network, the performance of object detection has been greatly improved. State-of-the-art object detection methods can be broadly classified into two categories, namely, one-stage and two-stage methods. The representative one-stage detectors include YOLO (which is an acronym for You Only Look Once) [4, 8–11], single-shot detector (SSD) [12], RetinaNet [13], and CenterNet [5, 6], which methods can perform nearly real-time detection, do not need proposal generation procedure, and directly conduct object detection in images. The YOLO families achieve state-of-the-art performance by integrating bounding boxes and subsequent feature resampling in a single stage. RetinaNet [13] can alleviate the fore-back class imbalance problem by Focal Loss. RefineDet [14] introduces a module to refine anchor boxes. CornerNet [15] proposes a method to eliminate anchor boxes, and an object is detected as a pair of keypoints (the top-left corner



(a) Yolov5 **(b)** CenterNet **(c)** GLE-Net

Fig. 1 Examples detection results of Yolov5, CenterNet, and our proposed method (GLE-Net). The red boxes represent undetected objects. Yolov5 algorithm, which undetected many people and bicycles. CenterNet can detect these small objects, whereas it failed to distinguish between people and bicycles

and bottom-right corner) of a bounding box. In CenterNet [6], an object is detected as one center keypoint and two keypoints of a bounding box, which contains the center location and other attributes of an object (e.g. size). In contrast, the most representative two-stage detectors, such as the R-CNN series and its variants. R-CNN [16–18] is one of the earliest and effective methods that adopt the deep convolutional neural network (CNN) for object detection, which replaces the traditional hand-crafted feature extracting process with CNN-based feature learning and improves the accuracy of object detection. There are two steps in two-stage detectors. In the first stage, focusing on generating a series of candidate region proposals that may contain objects. In the second stage, feature maps are extracted by region-of-interest (ROI) pooling from each proposal for classification and localization tasks. Fast R-CNN [17] generates region proposals on the feature map rather than the original input images, which improve detection efficiency by a large margin. Faster R-CNN [16] introduces an RPN to generate region proposals from the convolutional neural network and achieves end-to-end calculation of object recognition. R-FCN [19] uses the full convolution network ResNet to replace VGG to improve the effect of feature extraction and classification. Cascade R-CNN [20] proposes multiple repeated networks and they are connected sequentially, which can increase the number of high IoU score samples and allow the detector to obtain well performance. You only look one-level feature (YOLOF) [21] proposes diluted encoder and uniform matching to optimize detection. In CE-FPN [22], the authors were inspired by sub-pixel convolution, and then proposes a sub-pixel skip fusion method to perform both channel enhancement and up-sampling. BorderDet [23] proposes an efficient BorderAlign to extract border features from the extreme point of the border to enhance the point feature. CvT-ASSD [24] modified transformer backbone module by adding the convolutional token embedding and convolutional projection into transformer encoder block, along with the multi-stage design of the network by convolutions and making this maintaining certain computational efficiency.

2.2 Aerial Image Object Detection

Along with the publication of a few large-scale annotated datasets, such as DOTA [25], VisDrone [7], and DIOR [26] for object detection in aerial images, lots of researchers have attempted to transfer detectors for natural images to aerial image object detection. RICNN [27] adds a method to learn the rotation invariant neural network model based on existing R-CNN architecture, which is used for multiple classifications arbitrary orientation object detection. ROI Transformer [28] designs a rotated ROI learner to transform a horizontal ROI into a rotated ROI. In addition, this network is based on the RROIs to propose RPS-ROI-Align to extract rotation-invariant features. In LEVIR [29], the authors propose a new adaptive updating method for object detection inference in aerial images under the condition of prior small objects. DFL-CNN [30] proposes a double focal loss convolutional neural network framework for aerial vehicle detection. A Context-Aware Detection Network (CAD-Net) [1], which learns global and local contexts of objects by capturing their correlations with the global scene and the local neighboring objects or features, respectively. The rotational region CNN (R^2CNN) [2] proposes a modification of Faster R-CNN to extract pooled features of bounding boxes with different pooled sizes and then detect arbitrarily oriented objects. The small, cluttered, and rotated object detector (SCRDet) [31] fuses multi-layer features with effective anchor sampling and adds a supervised pixel attention network and channel attention network for small object detection. Furthermore, detecting small, cluttered, and rotated objects detector (SCRDet++) [32] devise an instance-level denoising module in the feature map for robust detection. The feature-merged single-shot detection (FMSSD) [33] aggregates the context information both in multiple scales and the same scale feature maps. SyNet [34] introduces a method multi-stage and single-stage in high-resolution aerial images to decrease the false-negative rate in multi-stage and increase the probability of proposals in single-stage. The multi-head rotated object detector (MRDet) [35] proposes an arbitrary-oriented region transformed from horizontal anchors to increase the original RPN and obtain accurate bounding boxes. R3Det [36] introduces an end-to-end refined one-stage rotation detector using a progressive regression approach from coarse to fine granularity.

3 The Proposed Method

In this section, we will describe the overall structure of our proposed method, shown in Fig. 2, and then explain the global and local ensemble network in detail. In this paper, we propose a global and local ensemble network to promote object detection accuracy. Specifically, in the global module, we first use a collection strategy to combine total detection results from multiply models and setting a dictionary T to store these results. Next, according to whether the object categories are consistent, the candidate bounding boxes in the dictionary T are saved in a list L. Furthermore, the bounding box with the highest confidence score is selected as the top priority box, which is used to match the remaining bounding boxes. We then continue to find the predictions with IoU greater than 0.50 as a subset M. According to our initial observations, the prediction boxes from various models should have assigned different weights. The specific statement formula refers to Eq. (9). In the local module, we normalize these confidence scores of all bounding boxes in

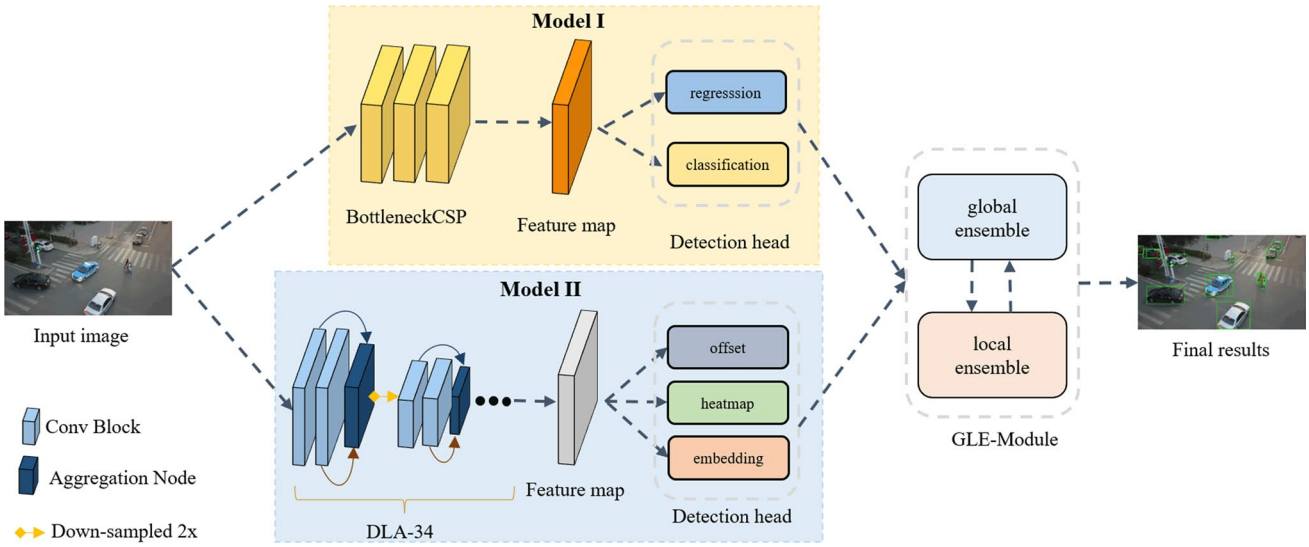


Fig. 2 Overview of our proposed framework (GLE-Net). Model I is Yolov5 and Model II is CenterNet with DLA-34 backbone. Firstly, the aerial images are fed into two different detectors, respectively. The top row (Model I) adopts BottleneckCSP (yellow cubic) as the backbone network to generate feature map automatically (orange cubic) and then predicts candidate bounding boxes by classification and regression module. The bottom row (Model II) adopts the deep layer aggregation with 34 convolutional layers to extract features (gray cubic) and then predicts heatmap, embedding and offset for the candidate objects. Finally, the global ensemble and the local ensemble module are used to merge all of the bounding boxes of those two models and output the final predicts

this subset M to obtain a new score of each bounding box, denoted as s_j. Finally, using the followed formulations to calculate a new optimization predict box, defined by Eq. (11).

3.1 The Model I: Yolov5

3.1.1 Backbone

In this work, we use three Bottleneck-CSPs to generate proposals and extract foreground object features from multi-layers using ROI-Align [37]. The backbone network is a local cross-layer fusion method to reduce the problem of excessive memory consumption. In addition, we apply standard data augmentation techniques that have proven effective for object detection, such as flipping, rotating and mosaic. Specifically, mosaic represents a new data augmentation approach that mixes four training images, which allows object detection outside their normal context and improves the accuracy of detection.

3.1.2 Detection Head

To effectively reduce the computational cost, we apply Bottleneck-CSPs to obtain the object feature map in the backbone network. Next, we use a similar top-down approach to collect feature maps from different stages. Besides, the spatial features and contextual features of the neck are extracted and fused to realize accurate object detection. Following the setting of the Yolo framework, at each scale, we predict three bounding boxes for each of the class-specific features maps. In addition, this model has always combined the classification and bounding box regression processes.

3.1.3 Loss Function

In this part, we will introduce the first model's loss function, which is composed of three losses formally defined as Eq. (1):

$$L_{\text{Model I}} = L_{\text{cls}} + L_{\text{obj}} + L_{\text{box}}. \tag{1}$$

The object confidence loss L_{obj} is binary cross-entropy loss and the classification loss L_{cls} is soft-max cross-entropy loss, shown as Eqs. (2) and (3). The bounding boxes loss L_{box} is L1-smooth loss, shown as Eq. (4). The specific expressions of the above three formulas are as follows:

$$L_{\text{obj}} = \begin{cases} \lambda_{\text{obj}-0} * \sum_{i=0}^{A^2} \sum_{j=0}^{B} 1_{i,j}^{\text{obj}-0} (c_{\text{pre}} - c_{gt})^2, \lambda_{\text{obj}-0} = 1 \\ \lambda_{\text{obj}-1} * \sum_{i=0}^{A^2} \sum_{j=0}^{B} 1_{i,j}^{\text{obj}-1} (c_{\text{pre}} - c_{gt})^2, \lambda_{\text{obj}-1} = 1 \end{cases}. \tag{2}$$

$$L_{\text{cls}} = \lambda_{\text{cls}} * \sum_{i=0}^{A^2} \sum_{j=0}^{B} 1_{i,j}^{\text{obj}-1} \sum_{c \in \text{class}} p_{\text{pre}_i}(c) \log(p_{gt_i}(c)). \tag{3}$$

$$L_{\text{box}} = \lambda_{\text{box}} * \sum_{i=0}^{A^2} \sum_{j=0}^{B} 1_{i,j}^{\text{obj}-1} * \text{GIoU}. \tag{4}$$

where A, B denote grid-scale and bounding box. $\lambda_{\text{obj}-0}$ and $\lambda_{\text{obj}-1}$ are weights of objectness loss and none-objectness loss, respectively. (i, j) represents the center point coordinate of a bounding box.λ_{cls} and λ_{box} are classification and regression hyperparameters. c is the true category of the object. $p(c)$ is the confidence score of the category c.$1_{i,j}^{\text{obj}-0}$ and $1_{i,j}^{\text{obj}-1}$ are the indicator functions. What's more, GIoU is optimized based on IoU, which not only focuses on overlapping areas but focuses on other non-coincident areas. Therefore, the above approaches can improve performance in object detection benchmarks.

3.2 Model II: CenterNet

3.2.1 Backbone

CenterNet has 4 architectures: ResNet-18 [38], ResNet-101 [38], DLA-34 [39], and Hourglass-104 [40]. In our experiments, we use CenterNet with a deep layer aggregation (DLA-34) backbone as another ensemble model for the task of detecting objects from an aerial image, where 34 represents 34 convolutional layers. This backbone is an image classification network with hierarchical skip connections, which utilize the full convolutional upsampling version of DLA for dense prediction and use iterative deep aggregation to increase feature map resolution symmetrically. In addition, we use deformable convolution to skip connections from the lower layer to the output layer, so that more object features in the aerial image can be preserved.

3.2.2 Detection Head

Object detection task could be treat as a keypoint estimation problem. CenterNet uses a center point of its bounding box to locate the object. And, this method uses keypoint estimation to find center points and regresses to all other object properties (e.g., size, orientation, 3D location and pose). Besides, CenterNet can simply extract a single center point per object without the need for grouping and post-processing and reduce negative bounding boxes. The detection head of Model II to predict heatmap, embedding and offset for the object. Heatmap is used to identify the heatmap of corners at the resolution of the feature map, embedding is applied to distinguish which corners belong to the same object, and offset is used to slightly adjust the l the locations of object on the heatmap.

3.2.3 Loss Function

The overall loss function (Eq. (5)) of the second network is defined as follows:

$$L_{\text{Model II}} = L_{\text{hm}} + \lambda_{\text{S}} * L_{\text{S}} + \lambda_{\text{O}} * L_{\text{O}}. \tag{5}$$

where L_{hm}, L_{S} and L_{O} are the heatmap loss, the size loss, and the offset loss of the prediction box, respectively. The hyperparameters λ_{s}, λ_{o} control the tradeoff and we set $\lambda_{\text{s}} = 0.1$ and $\lambda_{\text{o}} = 0.1$. In addition, L_{hm} is similar to focal loss. L_{S}, L_{O} both are mean absolute error (L1 loss). The specific expressions of the above three losses are defined as Eqs. (6)–(8).

$$L_{\text{hm}} = -\frac{1}{N} \sum_{xyc} \begin{cases} \left(1 - Y_{xyc}^{\text{pred}}\right)^2 \log\left(Y_{xyc}^{\text{pred}}\right), & \text{if } Y_{xyc}^{\text{gt}} = 1 \\ \left(1 - Y_{xyc}^{gt}\right)^4 \left(Y_{xyc}^{\text{pred}}\right)^2 \log\left(1 - Y_{xyc}^{\text{pred}}\right), & \text{otherwise} \end{cases}. \tag{6}$$

$$L_{\text{S}} = \frac{1}{N} \sum_{n} \left| S^{\text{pred}} - S^{\text{gt}} \right|. \tag{7}$$

$$L_{\text{o}} = \frac{1}{N} \sum_{gt_i} \left| O^{\text{pred}} - O^{\text{off}} \right|. \tag{8}$$

where N represents the number of keypoint in the image. $Y_{xyc}^{\text{pred}} \in [0,\ 1]^{\frac{H}{R} * \frac{W}{R} * C}$, R represents the stride of the output image relative to the original, C denotes the number of categories. W and H represent the width and height of images. $S^{\text{pred}} \in R^{\frac{H}{R} * \frac{W}{R} * 2}$ is the network output result. S^{gt} represents the width and height of a predicted box. gt_i represent the center point of the object box. $O^{\text{off}} = \frac{gt_i}{R} - \frac{gt_i}{R}$ and $O^{\text{pred}} \in R^{\frac{H}{R} * \frac{W}{R} * 2}$ is the backbone output offset value.

3.3 Global and Local Ensemble Module

In this part, we have bounding box predictions for the same image from N various models. In this paper, we choose Yolov5 and CenterNet as our benchmarks. The characteristics and structure of these networks have been introduced above in Sects. 3.1 and 3.2.

The global and local ensemble module includes two parts, namely the global section and the local section. In the global section, these works in the following steps:

(i) Traverse all the prediction results of each model. According to the image ID, these predictions bounding boxes into the dictionary T, and then return this dictionary T. Specifically, image ID is the key of the dictionary, and the value of the dictionary is composed of model ID, coordinates of the box, score, and category.

(ii) According to this dictionary T, each image ID is traversed. For all predicted objects in the image ID, they are classified based on dataset categories. Objects belonging to the same category are stored in a list L_i $(i = 1,\ 2,\ 3, \ldots, n)$, where i denotes the sequence number of categories.

(iii) In this category, the list is sorted in descending order of the classification confidence scores C.

(iv) Select a predicted bounding box with the largest confidence score in each list L_i $(i = 1,\ 2,\ 3, \ldots, n)$, as the top priority candidate box, and denoted as $B^1_{L_i}$.

(v) Declare a new subset M for the matchboxes. Use $B^1_{L_i}$ to iterate through the remaining predicted boxes in L_i and try to find the matching boxes. The match criterion is defined as a subset M, which is composed of the top priority candidate box $B^1_{L_i}$ and the rest prediction boxes with intersection-over-union (IoU) greater than 0.50. The IoU is formally defined as $\text{IoU} = \frac{\text{area}\ (b^{\text{p}} \cap b^{\text{g}})}{\text{area}\ (b^{\text{p}} \cup b^{\text{g}})}$, where b^{p} and b^{g} represent predicted and ground-truth bounding boxes, respectively.

(vi) According to the original detection results of different models, the same object of different models is given various weights. We propose a formulation (Eq. (9)) for learning the weight of the prediction boxes.

$$ws_i = \begin{cases} 1_{\text{box}_i \in m_1}\ (s_i * \lambda_1) \\ 1_{\text{box}_i \in m_2}\ (s_i * \lambda_2) \\ 1_{\text{box}_i \in m_1 \text{ and } m2}\ (s_i * \lambda_3) \end{cases} \tag{9}$$

where box_i denotes the ith prediction bounding box. m_1 and m_2 represent the first model and the second model. s_i $(i = 1,\ 2,\ 3\ldots)$ indicates the confidence of the ith bounding box. 1 denotes the indicator function. In this paper, $\lambda_1,\ \lambda_2,\ \lambda_3$ are hyper-parameters to balance the weight of each box, and we set $\lambda_1 = 0.9$, $\lambda_2 = 1.1$ and $\lambda_3 = 1.3$.

In addition, the calculation method for the local section is as follows:

(vii) Normalize the confidence scores of all predicted boxes in this subset M to get the new confidence scores of their respective prediction boxes, denoted as ws_i $(i = 1,\ 2,\ 3\ldots, n)$. And the normalization formula is expressed as:

$$ws_i = e^{ws_i} / \sum_n e^{ws_n}. \tag{10}$$

(viii) Use the following equation to calculate an optimization box as follows:

$$\begin{cases} x_{\text{new}} = x_1 * ws_1 + \cdots + x_k * ws_k \\ y_{\text{new}} = y_1 * ws_1 + \cdots + y_k * ws_k \\ w_{\text{new}} = w_1 * ws_1 + \cdots + w_k * ws_k \\ h_{\text{new}} = h_1 * ws_1 + \cdots + h_k * ws_k \end{cases}. \tag{11}$$

where $(x_{\text{new}},\ y_{\text{new}},\ w_{\text{new}},\ h_{\text{new}})$ and $(x_k,\ y_k,\ w_k,\ h_k)$ $(k = 1, 2, 3\ldots)$ represent an optimization box and the coordinates, width, and height of the center point of the original prediction box. The number k corresponds to the candidate box included in the list L_i, and ws_i represents the new score of each prediction box calculated in the sixth step. The score of the optimization box is replaced with the highest confidence score.

4 Experiments

4.1 Datasets and Metric

4.1.1 Aerial Image Dataset

VisDrone2019 [7] is a large-scale visual object detection benchmark, which was collected by Tianjin University. The VisDrone2019 DET [7] dataset for aerial object detection consists of 6471 aerial images for training and 548 images for the test, which were taken by camera-equipped unmanned air vehicles. The dataset annotated contains 10 object classifications: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. Each image scale ranges from 540×960 to 2000×1500 pixels and contains various shapes and scales. Since the aerial image detection task is still challenging because of class imbalance and object-image size mismatch, this dataset is utilized in this work for the validation of our proposed method.

4.1.2 Evaluation Metrics

The evaluation standard adopted in this paper is the mean average precision (mAP) in MS COCO [41], which is utilized to evaluate the performance of our method relative to other benchmarks. We computed three different average precision metrics: AP_{50}, AP_{75} and mAP. For AP_{50} and AP_{75} both consider a bounding box prediction as true, and overall object categories when the interest of union (IoU) scores between the predicted and the ground-truth bounding box must be larger than 0.5 and 0.75, respectively. The mAP, which takes a value between 0 and 1, is the average of all 10 IoU thresholds from a range of [0.5, 0.95] with a step size of 0.05.

Table 1 The detection performance on the VisDrone2019 validation dataset

Method	mAP	AP_{50}	AP_{75}
CornerNet [15]	17.41	34.12	15.78
Yolov3 [10]	13.8	30.43	11.18
CenterNet [43]	14.2	19.3	15.5
RefineDet512 [44]	14.9	28.76	14.08
Light-RCNN [45]	16.53	32.78	15.13
FPN [46]	16.51	32.2	14.91
Cascade-RCNN [20]	16.09	16.09	15.01
FRCNN + FPN [16]	21.4	**40.7**	19.9
CenterNet*	16.9	32.1	15.5
Yolov5	22.1	36.2	22.6
GLE-Net	**23.1**	39.0	**23.5**

The * denotes [5]. The bold numbers indicate the highest values in each column

4.2 Experimental Details

We use bottleneck-CSPs and DLA-34 as the backbones for our detection structure, and both have been pre-trained on the ImageNet [42]. Our proposed framework is shown in Fig. 2. In the training and testing stage, the input images are resizing to 608×608. In the training phase, we trained the model for 300 epochs with one batch size of 6 and a learning rate of 0.001. We have implemented the proposed method on PyTorch 1.5.0 and trained it based on Yolov5 and CenterNet. Our proposed model is continued to be trained on one server with an NVIDIA GeForce GTX 2080Ti GPU. In this experiment, we modified the number of Yolov5 output, in which only 100 boxes were selected as candidate boxes for each object. This is consistent with the number of candidate boxes by CenterNet.

4.3 Analysis of Comparison Results

To demonstrate the effectiveness of GLE-Net, we compared our model with the SOTA detection methods on the VisDrone2019 validation dataset in Table 1. We use CenterNet and Yolov5 in our proposed method. Compared to CornerNet, Yolov3, RefineDet512, Cascade RCNN, and Faster RCNN, GLE-Net achieves the best performance of 23.1% mAP with global and local ensemble strategy. Compare to original Yolov5 with the same backbone, GLE-Net improves the AP by 6.2% (from 16.9 to 23.1%). GLE-Net also outperforms the original CenterNet with DLA-34 backbone by 1.0% (from 22.1 to 23.1%). Specifically, GLE-Net improves nearly 8% points compared with the original CenterNet and exceeds 0.9% points with Yolov5 in terms of AP_{75} which indicates the flexibility and robustness of GLE-Net at higher IoU thresholds. However, the result of AP_{50} does not surpass FRCNN + FPN. The possible reason for this condition is that FRCNN + FPN is a two-stage algorithm, which have been proved better than one-stage algorithm in most cases. Nevertheless, our proposed method performance on AP_{50} goes beyond one-stage methods (CenterNet and Yolov5).

GLE-Net proposes to increase the accuracy of the regression box using global and local ensemble modules. The idea of the ensemble is also used by object detection competition or machine learning methods, such as IEEE Global Road Damage Detection Challenge. Therefore, to further improve the performance of aerial object detection, we also use global and local ensemble modules. Table 2 shows that the GLE-Net helps improve the performance from 16.9% and 22.1% to 23.1%, especially for all categories with small objects. Specifically, the improvements for car, pedestrian, van, truck, and bus are 9.3%, 6.1%, 5.1%, 9.5% and 13.2%, respectively. To verify the effectiveness of our method, a set of experiments was also done.

Figure 3 depicts the relationship between recall and precision curves of Yolov5, CenterNet, and GLE-Net algorithms in the VisDrone2019 dataset. It is obvious that the anchor-based method (Yolov5 in the green curve) is significantly better than the anchor-free method (CenterNet in the red curve). In the relationship between recall-precision curves, our GLE-Net method also performs better than the above methods. Specifically, we can see that the detection effect of the GLE-Net algorithm is better in the four categories of pedestrian, car, truck, and bus.

4.4 Experimental Results on VisDrone

The aerial images often contain small, dense objects in some regions. As shown in Fig. 4, when analyzing a straight road, the object nearly an aerial camera is larger and the far is smaller. In addition, for some objects, the edges of

Table 2 Detection results on the VisDrone2019 validation dataset

Method	mAP	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor
CenterNet*	16.9	13.1	9.7	3.9	43.8	23.7	16.6	10.3	6.5	28.1	13.3
Yolov5	22.1	17.8	11.2	6.2	51.5	27.3	25.7	14.4	9.1	40.5	17.5
GLE-Net	**23.1**	**19.2**	**12.4**	**6.9**	**53.1**	**28.8**	**26.1**	**15.2**	**10.0**	**41.3**	**18.4**

The * denotes [5]. The bold numbers denote the highest values in each column

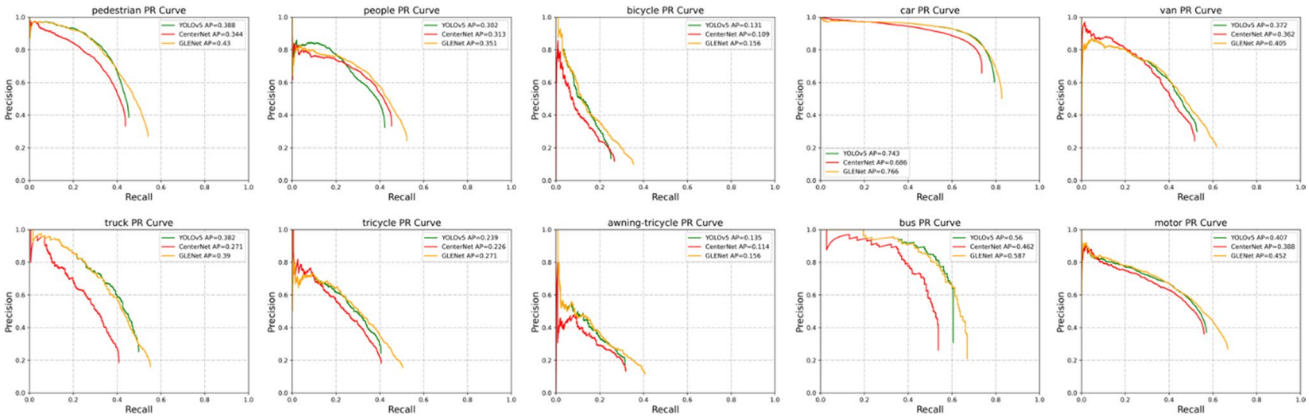


Fig. 3 The relationship between precision and recall curves of Yolov5, CenterNet, and GLE-Net in the VisDrone2019 dataset, respectively. The yellow lines represent GLE-Net, the red lines represent CenterNet and the green lines denote Yolov5, respectively

(a) Yolov5 **(b)** CenterNet **(c)** GLE-Net

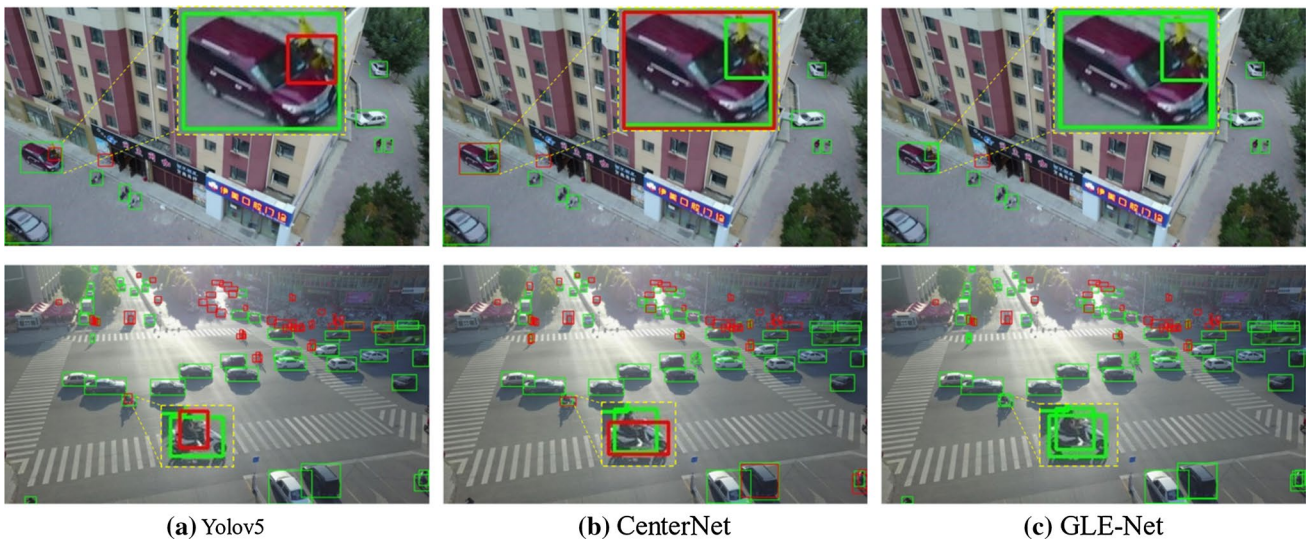
Fig. 4 Visualization of three detectors for the night scene in aerial images. The red boxes denote undetected objects, and the green boxes denote detected objects

foreground features are not very different from the background features, and the boundaries are blurry in the night scene. As shown in Fig. 4a, Yolov5 adopts a multi-scale prediction strategy, so that it can fuse the image features of edge information with different scales to obtain better details. Therefore, even when both the foreground and the background are influenced blurred, and Yolov5 can still detect edge objects. Unfortunately, the multi-scale features are extracted from Yolov5 is a low-resolution map that will obtain much lower accuracy with many dense and tiny missing objects. In contrast, CenterNet applies DLA-34 as the backbone network. Its characteristic is that all input and output feature maps have the same large spatial resolution, and they can express more small object features. Thus, CenterNet performs better than Yolov5 on overlapping small objects. However, for the foreground and background information not easily distinguishable, CenterNet does not perform well. Fortunately, GLE-Net successfully extracts the small objects and the edge objects. These results indicate that the global and local ensemble strategy of this paper combines the advantages of these two models, which can effectively reduce the missed detection rate.

Aerial images taken by unmanned aerial vehicles will contain many dense small objects, most of which are overlapping or unevenly distributed. In addition, there will be different aspect ratios of objects at different heights and angles in aerial images. From the overall results of Fig. 5, the proposed GLE-Net is substantial to Yolov5 and CenterNet. The reason is that the global and local ensemble approach by adding an appropriate weight for the detected

(a) Yolov5 **(b)** CenterNet **(c)** GLE-Net

Fig. 5 Visualization of three detectors for different roads in aerial images. The red boxes denote undetected objects, and the green boxes denote detected objects

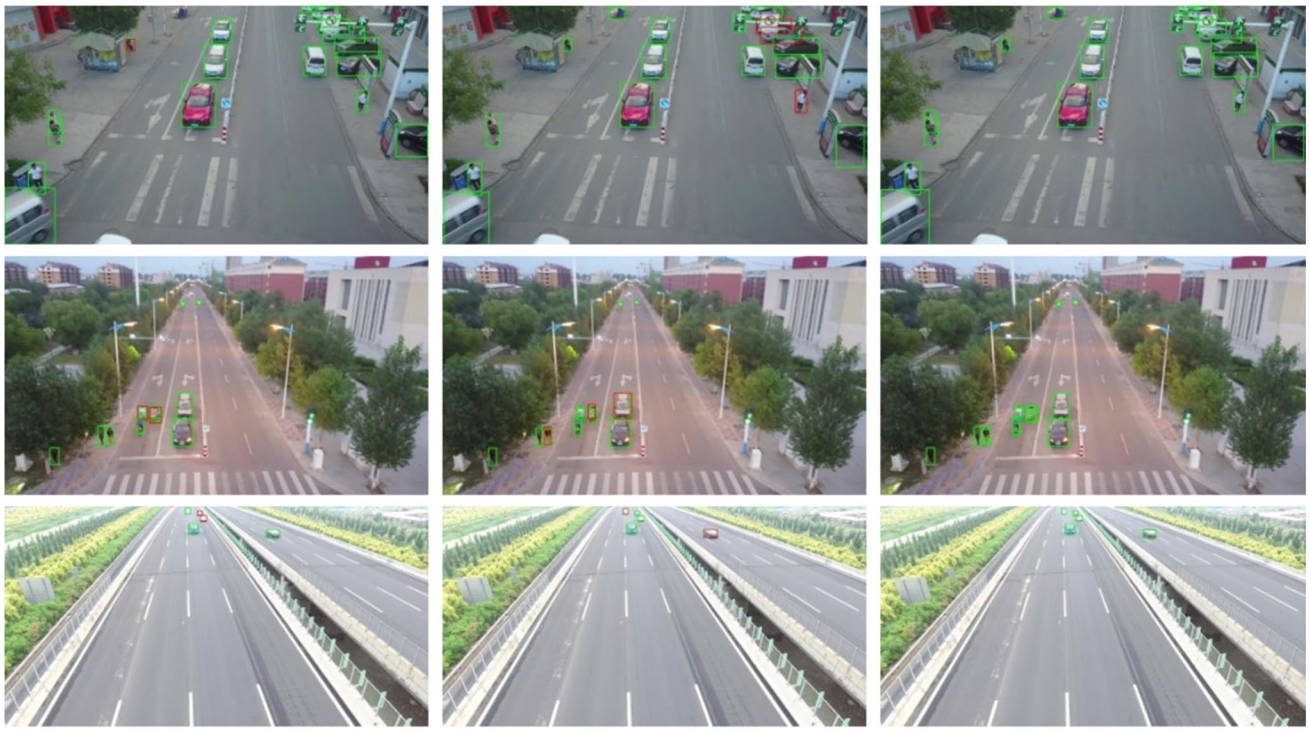


(a) Yolov5 **(b)** CenterNet **(c)** GLE-Net

Fig. 6 Visualization of three detectors for closing two objects in aerial images. The red boxes denote undetected objects, and the green boxes denote detected objects

objects so that reducing the undetected rate and enhancing the detection accuracy.

Visualization results of aerial images between close two objects are shown in Fig. 6. In the first row, a car and a bicycle are close together and the car partially obscures the rear of the bicycle, only the front of the bicycle can be seen. We can see that Yolov5 detects the car, but the bicycle behind it is not. The possible reason is that Yolov5 cannot identify incomplete objects, which leads to missed detection. Whereas CenterNet is the opposite, and we can observe that the corners of the bicycle and the car overlap, and these two objects may be identified as one object, so the car cannot be detected. Our algorithm can combine the advantages of the two methods to detect these objects. Similarly, the advantages of our approach are also shown in the second row. Some features of people and bicycles are weakened under strong light. In Yolov5, the related features of people cannot be extracted, and then the people cannot be detected. In CenterNet, the corner of the candidate box of people and bicycle features are close to being separated, which leads to the incorrect recognition of the two objects. However, in our proposed method, the advantages of the two methods are used to obtain the final detection result.



(a) Yolov5 **(b)** CenterNet **(c)** GLE-Net

Fig. 7 Qualitative results of Yolov5, CenterNet, and GLE-Net in aerial images. The red boxes denote undetected objects, and the green boxes denote detected objects

4.5 Qualitative Results

Figure 7 shows a few sample images from the VisDrone2019 dataset and the corresponding detection using the baseline model: Yolov5 and CenterNet (in the first column) and the proposed GLE-Net (in the second column). Qualitative results show that our proposed GLE-Net combines undetected objects in different situations, and then increases the detection precision of both algorithms. As Fig. 7 shows, the SOTA generic detection technique Yolov5 tends to produce miss detection under different road scenarios such as people in the first two examples and cars in the third example. In addition, the anchor-based method CenterNet undetected detection because of inaccurate corner matching. As a comparison, the proposed GLE-Net is capable of correctly detecting those objects under various adverse scenarios as illustrated in the third column of Fig. 7. The outstanding detection performance is largely attributed to the inclusion of the global contexts and local contexts (as described in Section III) within the proposed GLE-Net.

4.6 Inference Time and Parameters

In this part, we compare the inference speed and parameters of the two baselines, as shown in Table 3. The parameters of Yolov5 and CenterNet are 89 MB and 74.99 MB, respectively. It can be seen that the parameters of the two models are not much different. In the inference speed, there is a difference. The reason is that the backbone network used by CenterNet is DLA-34 [39], which is a multi-layer combination that spans the entire network, and then the inference time of its model will increase.

Table 3 Inference time and parameters comparison with Yolov5 and CenterNet on VisDrone2019

Method	Params (MB)	inference time (ms)
Yolov5	89.0	6.6
CenterNet	74.99	28

4.7 Ablation Study

In this subsection, to demonstrate the effectiveness of the global and local ensemble network and show this plug-and-play approach, we adopt three network methods to evaluate our proposed strategy. As shown in Table 4, the mAPs of the three baseline methods are 16.9%, 20.9%, and 22.1%, respectively. In this experiment, we applied our strategy by

Table 4 Ablation study of GLE-Net on VisDrone2019 validation dataset

Method	mAP	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}
CenterNet	16.9	32.1	15.5	9.2	26.3	39.2
Yolov4	20.9	34.9	21.3	11.8	31.8	47.6
Yolov5	22.1	36.2	22.6	12.4	33.8	53.0
Yolov4 + Yolov5	23.9	38.9	24.2	**14.0**	36.1	**55.5**
Yolov4 + CenterNet	21.9	37.9	21.6	12.1	33.6	47.2
Yolov5 + CenterNet	23.1	39.0	23.4	13.0	35.1	52.2
Yolov4 + Yolov5 + CenterNet	**24.2**	**40.3**	**24.4**	13.9	**36.6**	55.1

The bold text denotes the top result

combining them in pairs. It can be find that the integrated modules have improved to a certain extent. Specifically, the ensemble of Yolov4 and Yolov5 can improve the mAP from 20.9 to 23.9%. Yolov4 + CenterNet and Yolov5 + CenterNet can improve mAP over the baselines by 5% and 6.2%. The integration of the prediction results of Yolov4, Yolov5 and CenterNet can achieve an accuracy of 24.2%, with an increase of 7.3%, 2.1% and 3.3% compared to their baselines, respectively. Experiment results show that our idea of constructing ensemble predictions is effective.

5 Conclusions

In this paper, we have presented a global and local ensemble network for objects in aerial images. Considering the advantages and disadvantages of two state-of-the-art object detection models (Yolov5 and CenterNet), an ensemble module with global and local object features was added. The module fused regression and classification information from different models, and that is independent of the underlying algorithm, which can serve as an efficient plug-and-play network to improve the detection accuracy of the arbitrary model. The experimental results on the VisDrone2019 dataset demonstrate the competitive results of our proposed method.

Acknowledgements This research was funded by the National Natural Science Foundation of China under Grant No. 41971424; the Natural Science Foundation of Fujian Province, China under Grant 2020J01701, in part by the Fujian Provincial Science and Technology Program Project under Grants JAT190318, and in part by the Scientific Research Foundation of Jimei University, China, under ZP2022008.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Zhang, G., Lu, S., Zhang, W.: CAD-Net: A context-aware detection network for objects in remote sensing imagery. Remote Sensing **99**, 1–10 (2019)
2. Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P., Luo, Z. R2CNN: Rotational region CNN for orientation robust scene text detection. arXiv preprint arXiv:1706.09579 (2017)
3. Casado-Garcıa, A., Heras, J.: Ensemble Methods for Object Detection, pp. 2688–2695. IOS Press, Amsterdam (2020)
4. Jocher, G., Nishimura, K., Mineeva, T., Vilariño, R.: YOLOv5 (2020)
5. Zhou, X., Wang, D., Krhenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
6. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6569–6578 (2019)
7. Du, D., Zhu, P., Wen, L., Bian, X., Liu, Z.M.: VisDrone-DET2019: the vision meets drone object detection in image challenge results. In: ICCV Visdrone Workshop (2019)
8. 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, pp. 779–788 (2016)
9. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger, pp. 6517–6525 (2017).
10. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767 (2018)
11. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: single shot MultiBox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, Cham (2016)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **99**, 2999–3007 (2017)

14. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4203–4212 (2018)
15. Law, H., Deng, J.: CornerNet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
17. Girshick, R. J. C. S. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
18. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
19. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
20. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
21. Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., Sun, J. You only look one-level feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13039–13048 (2021)
22. Luo, Y., Cao, X., Zhang, J., Guo, J., Shen, H., Wang, T., Feng, Q.: CE-FPN: enhancing channel information for object detection. arXiv preprint arXiv:2103.10643 (2021)
23. Qiu, H., Ma, Y., Li, Z., Liu, S., Sun, J.: Borderdet: border feature for dense object detection. In: European Conference on Computer Vision, pp. 549–564. Springer (2020)
24. Jin, W., Yu, H.J.: CvT-ASSD: convolutional vision-transformer based attentive single shot MultiBox detector. arXiv preprint arXiv:2110.12364 (2021)
25. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: a large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974–3983 (2018)
26. Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: a survey and a new benchmark. Remote Sensing **159**, 296–307 (2020)
27. Gong, C., Zhou, P., Han, J.: Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. Remote Sensing **54**(12), 7405–7415 (2016)
28. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning RoI transformer for oriented object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2849–2858 (2020)
29. Zou, Z., Shi, Z.: random access memories: a new paradigm for target detection in high resolution aerial remote sensing images. IEEE Trans. Image Process. **27**(3), 1100–1111 (2018)
30. Yang, M.Y., Liao, W., Li, X., Cao, Y., Rosenhahn, B.J.P.E.: Vehicle detection in aerial images. Remote Sensing **85**(4), 297–304 (2019)
31. Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Xian, S., Fu, K.: SCRDet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8232–8241 (2019)
32. Yang, X., Yan, J., Yang, X., Tang, J., Liao, W., He, T.: SCRDet++: detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. arXiv preprint arXiv:2004.13316 (2020)
33. Wang, P., Sun, X., Diao, W., Fu, K.: FMSSD: feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. Remote Sensing **58**(5), 3377–3390 (2019)
34. Albaba, B.M., Ozer, S. SyNet: an ensemble network for object detection in UAV images. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 10227–10234 (2021)
35. Qin, R., Liu, Q., Gao, G., Huang, D., Wang, Y. MRDet: a multi-head network for accurate oriented object detection in aerial images. arXiv preprint arXiv:2012.13135 (2020)
36. Yang, X., Liu, Q., Yan, J., Li, A., Zhang, Z., Yu, G.: R3det: refined single-stage detector with feature refinement for rotating object. arXiv preprint arXiv:1908.05612 (2019)
37. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
38. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
39. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2403–2412 (2018)
40. Newell, A., Yang, K., Jia, D.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, pp. 483–499. Springer, Cham (2016)
41. Belongie, S.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer, Cham (2014)
42. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
43. Tang, Z., Liu, X., Shen, G., Yang, B.: PENet: object detection using points estimation in aerial images. arXiv preprint arXiv:2001.08247 (2020)
44. Jadhav, A., Mukherjee, P., Kaushik, V., Lall, B.: Aerial multi-object tracking by detection using deep association networks. In: 2020 National Conference on Communications (NCC). IEEE, pp. 1–6 (2020)
45. Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., Sun, J.: Light-head r-cnn: In defense of two-stage object detector. arXiv preprint arXiv:1711.07264 (2017)
46. 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, pp. 2117–2125 (2017)