



Object Detection in Aerial Images Using a Multiscale Keypoint Detection Network

Jinhe Su , JiaJia Liao, Dujuan Gu, Zongyue Wang , and Guorong Cai

Abstract—Automatic object detection in aerial imagery is being increasingly adopted in many applications, such as traffic monitoring, smart cities, and disaster assistance. In keypoint-based detectors, the prediction modules are usually generated from a fixed feature map scale. This configuration significantly limits the ability to detect multiscale objects in aerial scenes. The corner selection module in these detectors often ignores that a category in an aerial image is relatively unitary. In this article, a novel network, called the multiscale keypoint detection network (MKD-Net), is proposed to address these challenges. MKD-Net fuses multiscale layers to generate multiple feature maps for objects of different sizes. During the inference phase, both feature maps can be exploited for predicting corners. Moreover, a category attention module is designed to reduce the channel noise for a single-category scene. Experiments on benchmarks PASCAL VOC and DOTA show promising performance of MKD-Net compared with the baseline network. The code is available on <https://github.com/jason-su/MKD-NET>.

Index Terms—Attention network, convolutional neural networks (CNNs), loss functions, object detection, unmanned aerial vehicles.

I. INTRODUCTION

OBJECT detection is one of the fundamental tasks in computer vision. Various deep convolutional neural network (CNN) based detectors have been proposed for natural images. Most of the current state-of-the-art methods use multiscale and multispect-ratio anchor boxes. Those methods regress them to the target size with supervision of ground-truth bounding boxes.

Many existing object detectors, such as faster region-based CNNs (R-CNN) [1], You only look once v3 (YOLOv3), and RetinaNet [2], have been widely used for aerial object detection. They rely on predefined anchor boxes. This scheme has two drawbacks: a very large number of anchor boxes are generated and several hyperparameters and design choices are introduced to satisfy objects with various sizes that typically need to be

carefully tuned in order to achieve good performance. Hence, anchor-free frameworks have drawn much attention. One of the most popular frameworks is keypoint detection, which aims at extracting the corners of objects, then generates bounding boxes as a pair of corners grouped together. CornerNet [3], the most representative framework, predicts the top-left and bottom-right corners of the bounding boxes, subsequently predicting the heatmaps of the corners with associated embeddings to group them. Nevertheless, the performance of CornerNet is still restricted by its relatively weak ability to generate heatmaps. Most of these keypoint-based detectors are vulnerable to problems of distinct, dense, small objects and group relevant paired key points under aerial scenes.

Compared with natural images, the sizes of objects in aerial images vary greatly. As shown in Fig. 1, in the dataset for Object detection in aerial images (DOTA) dataset, the areas of objects range from 12 pixels to 640 000 pixels. More than 70% of objects have between 256 and 4096 pixels. We observe that approximately 5% of the objects have an area of less than four pixels, and more than 1% of objects have an area of more than half of the feature map with an $8\times$ -reduced resolution compared to the original image. Scale variations have become a challenging task in the field of aerial object detection. Generally, it is unsuitable to generate key points based on a single-scale feature map for all kinds of object sizes. Feature pyramids are widely applied in anchor-based detectors, such as YOLOv3 [4], feature pyramid networks (FPNs), because they can extract and fuse rich semantics better from all levels. Thus, to obtain accurate detection in key point-based detectors, we design a multiscale fuse module to improve the feature extraction ability.

In CornerNet, there are two sets of multichannel heatmaps: one heatmap for top-left corners, the other one heatmap for bottom-right corners. Each channel of the heatmap corresponds to the corner of a category. During the test, CornerNet selects the top k corners from all the channels to generate bounding boxes. However, more than 65% of the images contained fewer than two types of objects in the DOTA dataset. As shown in Fig. 2, there are only planes on the tarmac, ships in the port, and small vehicles in the parking lot. Hence, an intuitive way to improve detection efficiency is to reduce the disturbance of irrelevant channels in feature maps. It can help the network better selects corners from these channels where there are objects in the corresponding category.

Inspired by these phenomena, we introduce the one-stage multiscale keypoint detection network (MKD-Net) detector for addressing the two aforementioned challenges, by adding a

Manuscript received September 24, 2020; revised November 30, 2020; accepted December 9, 2020. Date of publication December 14, 2020; date of current version January 6, 2021. This work was supported in part by the Natural Science Foundation of Fujian Province, China under Grant 2020J01701, in part by the Scientific Research Foundation of Jimei University, China, under Grant ZQ2019013, and in part by the Fujian Provincial Science and Technology Program Project under Grants JAT190318. (Corresponding author: Guorong Cai.)

Jinhe Su, JiaJia Liao, Zongyue Wang, and Guorong Cai are with the School of Computer Engineering, Jimei University, Xiamen 361021, China (e-mail: sujh@jmu.edu.cn; jiajialiao@jmu.edu.cn; wangzongyue@jmu.edu.cn; guorongcai.jmu@gmail.com).

Dujuan Gu is with NSFOCUS Information Technology Company, Ltd., Beijing 100000, China (e-mail: gudujuan@sina.com).

Digital Object Identifier 10.1109/JSTARS.2020.3044733

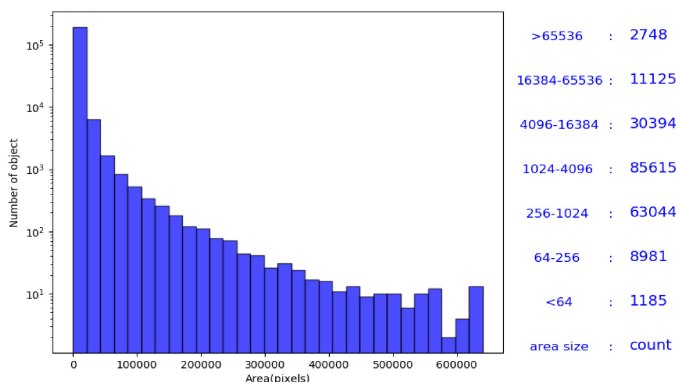


Fig. 1. Distribution of object sizes in the DOTA dataset.

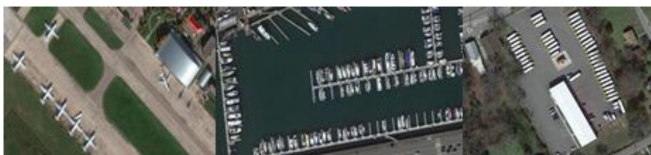


Fig. 2. Three sample images from the DOTA dataset.

multiscale network and a category attention network. MKD-Net can be considered a variant of CornerNet-squeeze [5] that detects an object bounding box as a pair of key points in one network. In MKD-Net, we add a multiconvolution module to fuse features from different layers. During inference, we make predictions from multiscale feature maps. We also add an attention network to predict the categories of objects in an image to reduce the noise in the background. In summary, the article has the following contributions.

- 1) We propose a novel network (MKD-Net) to simultaneously address the challenges of large size variations and small dense objects for object detection in aerial imagery.
- 2) We design an effective multiscale network to fuse powerful features from the backbone and select corners from the multiscale feature map.
- 3) We design a supervised attention network to reduce the adverse impact of category noise during testing.

II. RELATED WORK

Recently, exciting breakthroughs have been made in object detection using deep convolutional networks. Current mainstream detector frameworks can be roughly categorized into two main types: anchor-based detectors and anchor-free detectors.

Anchor-Based Detectors: Anchor-based detectors require generating a set of bounding boxes along with their labels. The most representative one is the R-CNN series and its variants [1], [6], [7], [25], [26], which are two-stage detectors. In the first stage, a region proposal network (RPN) that is based on the sliding-window mechanism generates a large number of candidate bounding boxes. In the second stage, feature maps are extracted by region-of-interest pooling from each bounding box for classification and bounding-box regression tasks. Li *et al.* [25] introduced a local-contextual feature fusion network based on Faster R-CNN. In contrast, the one-stage detectors,

such as the single-shot multibox detector [8], YOLOv3 [4], and RetinaNet [2], directly regress the bounding boxes, which leads to high efficiency while sacrificing accuracy. However, a very large set of bounding boxes can be generated by the anchor-based detector to ensure sufficient overlap with most ground-truth boxes. As a result, those methods will face a large class imbalance during training [2]. They also introduce several hyperparameters for the design of anchor boxes [3]. Generally, these anchor frames have two different representation methods. One is that the anchor frame contains only the upper-left corner and the lower-right corner, the other is to use the four points of the anchor frame. In particular, the more popular representation method in aerial imagery is to use four points to represent the anchor frame because the four points can represent not only a regular rectangular frame but also a quadrilateral frame or a diamond frame, which better represent the target position in an image.

Anchor-Free Detectors: With the advent of CornerNet [3] and CenterNet [9], which replace bounding-box supervision with keypoint supervision, anchor-free detectors have outperformed their anchor-based counterparts. The feature selective anchor-free module for single-shot object detection [10], guided anchoring [11], the fully convolutional one-stage object detector [12] and FoveaBox [13] replace the bounding boxes with anchor points and point-to-boundary distances. Those detectors have required much more complicated post-processing. They have been considered unsuitable for generic object detection due to difficulty in handling overlapping bounding boxes and a relatively low.

Object Detection in Aerial Images: Along with the publication of a few large-scale annotated datasets (i.e., DOTA [14], VisDrone [15], DIOR [28] and the Northwestern Polytechnical University ten-class geospatial object detection dataset (NWPU VHR-10) [16]) for object detection in aerial images, many studies have attempted to transfer detectors for natural images to aerial object detection. Deng *et al.* [17] proposed coupled R-CNNs for aerial vehicle detection. Sommer *et al.* [18] redesigned the anchor settings and backbone structure based on Faster R-CNN to detect vehicles in aerial images. The scale-adaptive proposal network [19] adds an RPN to shallow feature maps to detect small objects in aerial images. The rotational region CNN [20], a modification of Faster R-CNN, extracts pooled features of bounding boxes with different pooled sizes to detect arbitrarily oriented objects. Cheng *et al.* [27] proposed to learn a rotation-invariant CNN model based on R-CNN framework used for multiclass arbitrary orientation object detection. ClusDet [23] proposed a method to extract a large number of target slices using clustering, then performed target detection based on the relatively sparse targets under aerial photography. In addition, the small, cluttered and rotated object detector (SCRDet) [21] fuses multilayer features with effective anchor sampling, adds a supervised pixel attention network and channel attention network for small and cluttered object detection. In general, these aerial object detectors are modified based on anchor-based detectors, which have the previously mentioned drawbacks. In this article, we propose a multiscale keypoint aerial object detector based on CornerNet-Squeeze.

we design a supervised category attention module to better select the top k corners by coding explicit prior knowledge, as shown in Fig. 3. Specifically, in the category attention network, the feature map $F1$ passes through an MCL with different channels, and then a one-dimensional (1-D) vector is learned through a fully connected layer. Each value in the vector represents the scores of a category that appear in a test image. Then, we apply channel reduction operations for $C1$ and $C2$ by 1×1 convolutional layers and output three feature maps. The channel numbers for these feature maps are 15, 1, and 1. The feature map $H1$ and $H2$ have 15 channels, where each channel corresponds to the corner distribution of a category in an image. The softmax operation is performed on the vector V . Then the output is separately multiplied by $H1$ and $H2$. Finally, two new information feature maps $N1$ and $N2$ are generated. To train the network in this process, we adopt a supervised learning method. First, we can easily obtain a binary map as a label according to the ground truth. Second, we use the cross-entropy loss of the binary map and the 1-D vector as the category attention loss.

D. Loss Function

The loss of MKD-Net consists of the corner loss, offset loss, push loss, pull loss, and category loss, defined as follows:

$$\text{Loss} = \sum_{i=1}^m (\lambda_1 L_{\text{det}}^i + \lambda_2 L_{\text{off}}^i + \lambda_3 L_{\text{push}}^i + \lambda_4 L_{\text{pull}}^i) + \frac{\lambda_5}{N_{\text{category}}} \sum_{i=1}^{N_{\text{category}}} L_{\text{category}}(p_i, g t_i). \quad (1)$$

The hyperparameters λ_1 , λ_2 , and λ_3 control the tradeoff. We set $\lambda_1 = 0.8$, $\lambda_2 = 1$, $\lambda_3 = \lambda_4 = 0.1$, $\lambda_5 = 5$; m denotes the number of feature maps, which is 4 in our experiment.

The first component L_{det} is the focal loss of the predicted heatmaps, which are used to detect the top-left and right-bottom corners of the bounding boxes. The ground-truth heatmaps have been augmented with unnormalized Gaussian distribution, which could reduce the penalty given to negative locations within a radius of the positive location.

As for CornerNet paper, e_{tk} has been used to denote the embedding for the top-left corner of object k , e_{bk} stands for the bottom-right corner. We then use the ‘‘pull’’ loss to group the corners and the ‘‘push’’ loss to separate the corners

$$L_{\text{pull}} = \frac{1}{N} \sum_{k=1}^N [(e_{tk} - e_k)^2 + (e_{bk} - e_k)^2] \quad (2)$$

$$L_{\text{push}} = \frac{1}{N(N-1)} \sum_{k=1}^N \sum_{j=1, j \neq k}^N \max(0, \Delta - |e_k - e_j|) \quad (3)$$

where $e_k = (e_{tk} + e_{bk})/2$. The value of e_k is between 0 and 1. The maximum distance is 1 for two corners from different objects. Hence, we set Δ to be 1 in all our experiments. The pull loss L_{pull} determines whether a pair comprising the top-left corner and bottom-right corner is from the same bounding box or not. The push loss L_{push} aims at separating the corners from different bounding boxes.

Given a network with downsampling layers, the coordinate (x, y) in an image is been mapped to $(\lfloor \frac{x}{n} \rfloor, \lfloor \frac{y}{n} \rfloor)$ in feature maps.

We use smooth $L1$ function to define the offset loss L_{off} , since smooth $L1$ is widely used to slightly adjust the corner locations. The category loss L_{category} is the softmax cross-entropy loss, which is used to determine what kinds of objects exist in an image.

IV. EXPERIMENTS

This section is divided by subheadings and provides a concise and precise description of the experimental results, their interpretation and the experimental conclusions that can be drawn.

A. Experimental Setup

Tests are implemented in PyTorch on a server with two NVIDIA GeForce GTX 2080 Ti GPUs and 11 GB RAM. We perform experiments on DOTA and VOC datasets to verify the generalizability of our techniques.

- 1) *Evaluation Metrics*: The evaluation standard adopted in this article is the average precision (AP), which is used to evaluate the performance of our methods relative to other benchmarks. The AP, which takes a value between 0 and 1, is the average of all ten intersection over union (IoU) thresholds from a range of [0.50, 0.95] with a step size of 0.05. The IoU thresholds for calculating AP50 and AP75 are set to 0.5, and 0.75, respectively, for all the categories. The mean AP (mAP) of all the categories refers to the average value of the APs for each category.
- 2) *DOTA Dataset and Preprocessing*: The experiment performed in this article uses the DOTA-V1.0 dataset, which is a recently published large-scale open-access dataset for benchmarking object detection in remote sensing imagery. It is likely the largest and most diverse dataset for this task. It contains 2806 aerial images that were captured using different sensors and platforms where over 188 000 object instances were annotated using quadrilaterals. The images from DOTA are diverse in size, ground sample distance, sensor type, etc. The captured objects also exhibit rich variation in terms of scale, shape, and orientation. Fifteen categories of objects are annotated: plane (PL); baseball diamond (BD); bridge (BR); ground track field (GTF); small vehicle (SV); large vehicle (LV); ship (SH); tennis court (TC); basketball court (BC); storage tank (ST); soccer ball field (SBF); roundabout (RA); harbor (HA); swimming pool (SP); and helicopter (HC). There are two detection tasks for the DOTA dataset: horizontal bounding boxes and oriented bounding boxes (OBBs). This dataset is divided into three subsets for training (1/2), validation (1/6), and testing (1/3), where the ground truth of the test set is not publicly accessible.

Optical remote sensing images are often massive, e.g., the size of DOTA images is usually between 800×800 and 4000×4000 pixels, but can be up to 6000×6000 pixels. These images contain objects exhibiting a wide variety of scales, orientations, and shapes. Because feature extraction networks based on CNNs cannot be used directly to fit the hardware memory in the training stage. Hence, we crop images into patches of size 600×600 pixels with an overlap of 150 pixels among neighboring patches.

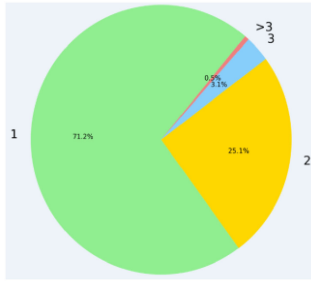


Fig. 4. Number of categories contained in one image.

Finally, 20 690 images were used for training and 5341 images were used for testing.

As shown in Fig. 4, after data preprocessing, 71.2% of images in the DOTA aerial image set contain only one category of objects, and 25.1% of images contain two categories of objects.

- 3) *PASCAL VOC Dataset*: We also evaluate our approach on PASCAL VOC 2007 and 2012 [29]. Both dataset consist of 20 categories. Pascal VOC 2007 contains a trainval set of 5011 images and a test set of 4952 images, while VOC 2012 consists of 11 540 images as trainval set and 10 991 as test set. For fair comparisons, all methods are trained on VOC 2007 trainval + VOC 2012 trainval, and tested on VOC 2007 testset.
- 4) *Training Details*: The backbone network is randomly initialized under the default setting of PyTorch with no pre-training on any external dataset. We adopt stochastic gradient descent with momentum for network optimization. No other data augmentation is performed. During training, we set the input resolution of the network to 511×511 , which leads to the output resolutions of $[32 \times 32, 64 \times 64]$, which correspond to $(C1, C2)$, respectively. Unless specified, the input images are resized to 511×511 . The batch size is set to 12 images per GPU. We train the network on two NVIDIA 2080 Ti GPUs with 11 GB memory. The model is trained with 200 000 iterations, which takes approximately 70 h on the DOTA dataset, and the learning rate changes during the 50 000 and 100 000 iterations from $5e-3$ to $5e-5$.
- 5) *Inference Details*: We forward the input image through the network to obtain two feature maps. Only the original image is used for testing, which means that the test data are not augmented. For each feature map, we pick top N corners from all channels in the top-left corner heat map and bottom-right corner heat map, respectively. The $2*N$ corners are paired off and generated $N*N$ candidate bounding boxes. We remove the boxes with low scores or position error, which is exactly the same operation in CornerNet-Squeeze [5]. Then, we apply soft-NMS [22] to suppress redundant detections to the overall bounding boxes. We finally choose the top k ($k = 100$ in our experiment) scores as the detection result from the remaining boxes.

Unlike CornerNet-squeeze, which uses only the features from the last layer of the whole network to make predictions, we use the features from two different layers to make predictions. To

avoid high overlaps in the prediction results between different layers, which may hurt the performance of the network, we further limit the number of top corners and the number of bounding boxes that can be generated by the layer.

B. Accuracy Evaluation

MKD-Net adds an extra layer based on CornerNet-Squeeze, which means that they share the same backbone network and have the same number of hyperparameters. Hence, we choose CornerNet-Squeeze as the baseline. In addition, we also compare our network with faster R-CNN, which is a two-stage object detection framework. For fairness, all the experimental hyperparameter settings are strictly consistent. The CornerNet-squeeze-128 and CornerNet-squeeze-128 multimodels are the variant of CornerNet-squeeze and CornerNet-squeeze multi that removing a downsampling layer before fed into the hourglass modules. Hence, the resolution of feature map $N2$ is $128*128$. Table I gives the performance of several models on the DOTA dataset. It can be seen that the mAP of our MKD-Net algorithm is higher than that of the other algorithms; the mAP of MKD-Net is 3.4% higher than that of CornerNet-squeeze and 9.7% higher than that of Faster R-CNN.

CornerNet-Squeeze multi and MKD-Net achieves much better performance than CornerNet-Squeeze, on the SH, ST, SV, and LV. As shown in Fig. 5, those categories of objects in images are primarily small and medium-sized. The large-size objects account for less 10% of each category. The performance improving for small-size objects might have been due to the extra feature map, $N2$ in Fig. 3, introduced by the multiscale module, which has larger resolution that could beneficial to small objects detect.

C. Ablation Study

We gradually add the multiscale module and category attention module to the baseline, CornerNet-squeeze, to investigate the effectiveness of the proposed MKD-net on DOTA dataset. We, first, apply the multiscale module to the head branch. As shown in the second row of Table II, the module leads to a gain of 2.8 on the AP on CornerNet-squeeze. We also see that the improvement mainly occurs in the AP with a low threshold. The improvement at a low IoU threshold is because the multiscale can increase the density of the bounding boxes and can potentially raise the chance of matching ground truth. This method could maintain the predictions with both a high classification score and localization.

Further study concerns the influence of category attention module. We address the concern that the category attention module might not provide a sufficiently good accuracy, and then leads to missed detection of objects in the same category. The third row in Table II gives that the detection performance was reduced slightly by adding the category attention module. The fourth row shows that adding the category attention module on the multiscale branch boosts the performance from 28.0 to 31.8 on CornerNet-squeeze. In our experiments, the detector achieves the best performance when using the multiscale module and category attention module.

TABLE I
PERFORMANCE EVALUATION ON THE DOTA DATASET. “MULTI” DENOTES A SQUEEZE NETWORK WITH A MULTISCALE MODULE AND INFERENCE WITH “N1, N2” FEATURE MAPS

Method	mAP	PL	SH	ST	BD	TC	BC	GTF	HA	BR	SV	LV	HC	RA	SBF	SP
Faster R-CNN	22.1	30.5	16.2	22.4	25.1	32.7	22.9	24.0	25.8	15.3	12.5	22.1	29.5	19.6	14.8	17.4
CornerNet-Squeeze	28.4	50.9	9.3	21.4	35.2	73.4	37.2	30.9	28.2	22.2	13.1	25.5	10.3	29.0	20.8	18.4
CornerNet-Squeeze-128	29.9	50.7	11.1	26.4	36.5	76.9	31.1	39.0	30.6	22.1	14.4	24.9	7.7	35.8	30.3	11.5
CornerNet-Squeeze multi	31.2	57.5	19.4	28.1	32.7	79.5	27.7	35.8	29.6	21.2	23.8	34.9	10.9	25.2	24.4	17.4
CornerNet-Squeeze-128 multi	31.7	54.0	18.9	32.5	34.5	78.8	30.7	35.9	30.4	22.1	23.4	34.3	9.5	28.7	27.4	15.1
MKD-Net	31.8	58.5	18.5	28.8	37.1	73.6	34.9	35.3	33.0	21.7	21.9	30.9	13.4	25.6	27.9	16.4

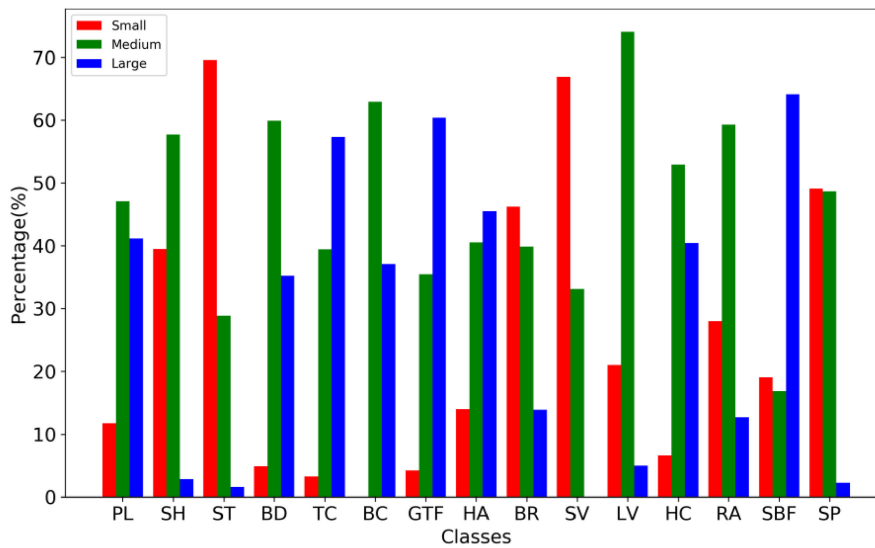


Fig 5. Proportion of small- (area<32*32), medium- (32*32<area<96*96), and large-sized (area>96*96) objects in each category.

TABLE II
ABLATION STUDIES ON THE MULTISCALE STRATEGY AND CATEGORY ATTENTION ON CORNERNET-SQUEEZE

multi-scale	category attention	CornerNet-Squeeze			
		AP	AP ₅₀	AP ₇₅	Params
		28.4	49.6	27.5	31.7 M
✓		31.2	54.6	30.0	41.2 M
	✓	28.0	46.0	28.0	37.3 M
✓	✓	31.8	53.8	31.7	44.0 M

In Table III, we also analyze the impact of the multiscale feature map during inference on DOTA dataset. To avoid the influence of the category attention network, we conduct experiments on CornerNet-Squeeze with a multiscale module. We use the feature maps “N2” and “N1, N2”, which are shown in Fig. 2, to make predictions. The feature map “N1” has half the resolution of feature map “N2.” Table III gives that the detection results of both methods have been improved for small and medium object detection after adding multiscale feature maps.

The mAP increases by 4%–7%. It is reasonable that the extra predicted layer costs approximately 8 ms of extra inference time due to the doubled candidate corners and the larger number of paired bounding boxes.

D. Evaluation on VOC Dataset

We also compare the performance on the PASCAL VOC dataset, as given in Table IV. It can be seen that the MKD-Net achieves the mAP of 44.8%, improving that of CornerNet-Squeeze by 4.5%. This comparison clearly suggests that our framework can also work better on natural scenes. However, although the method has achieved the better performance, the detection accuracy is still low. One of the possible reasons is the characteristics of VOC dataset. The VOC dataset have images that are not square. The resize operation before fed the images into network could change the appearance features. The other reason maybe that we focused on real time detector. We modified the network based on CornerNet_squeeze, a variant of CornerNet that sacrificing accuracy for inference speed. There are three methods, CenterNet, CornerNet and CornerNet Squeeze, that have similar network architectures. The accuracy

TABLE III
ABLATION STUDIES ON THE INFERENCE WITH MULTIPLE FEATURE MAPS ON CORNERNET-SQUEEZE. "N2" AND "N1, N2" DENOTE THE FEATURE MAPS USED TO MAKE THE PREDICTIONS

	CornerNet Squeeze multi					CornerNet Squeeze-128 multi				
	mAP	small	medium	large	time	mAP	small	medium	large	time
N2	29.0	12.1	33.7	30.3	27ms	30.3	12.4	35.3	28.6	57ms
N1, N2	31.2	14.6	35.9	30.6	33ms	31.7	14	36.6	29.2	66ms

TABLE IV
CATEGORY PERFORMANCE COMPARISONS ON PASCAL VOC DATASET

	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv
CornerNet Squeeze	40.3	52.0	44.9	32.0	26.3	16.2	59.4	48.7	60.9	17.2	41.5	32.1	47.7	49.2	38.4	43.6	14.7	39.8	43.0	58.5	40.7
MKD-Net	44.8	53.6	48.0	33.5	31.4	19.9	61.1	52.0	64.0	23.0	44.7	46.1	53.0	56.0	47.1	45.4	18.6	41.1	51.4	61.4	45.5

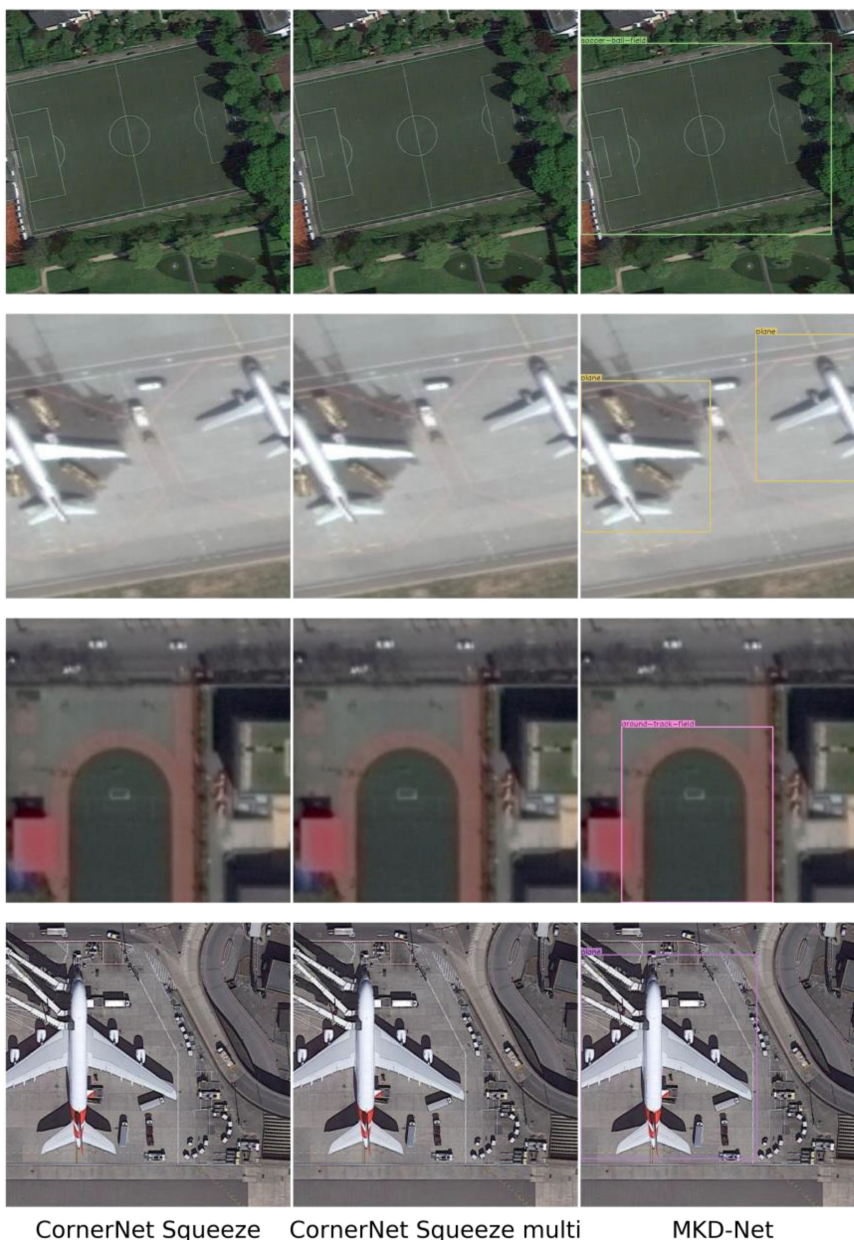


Fig 6. Example results of the three detectors for very large objects in images.

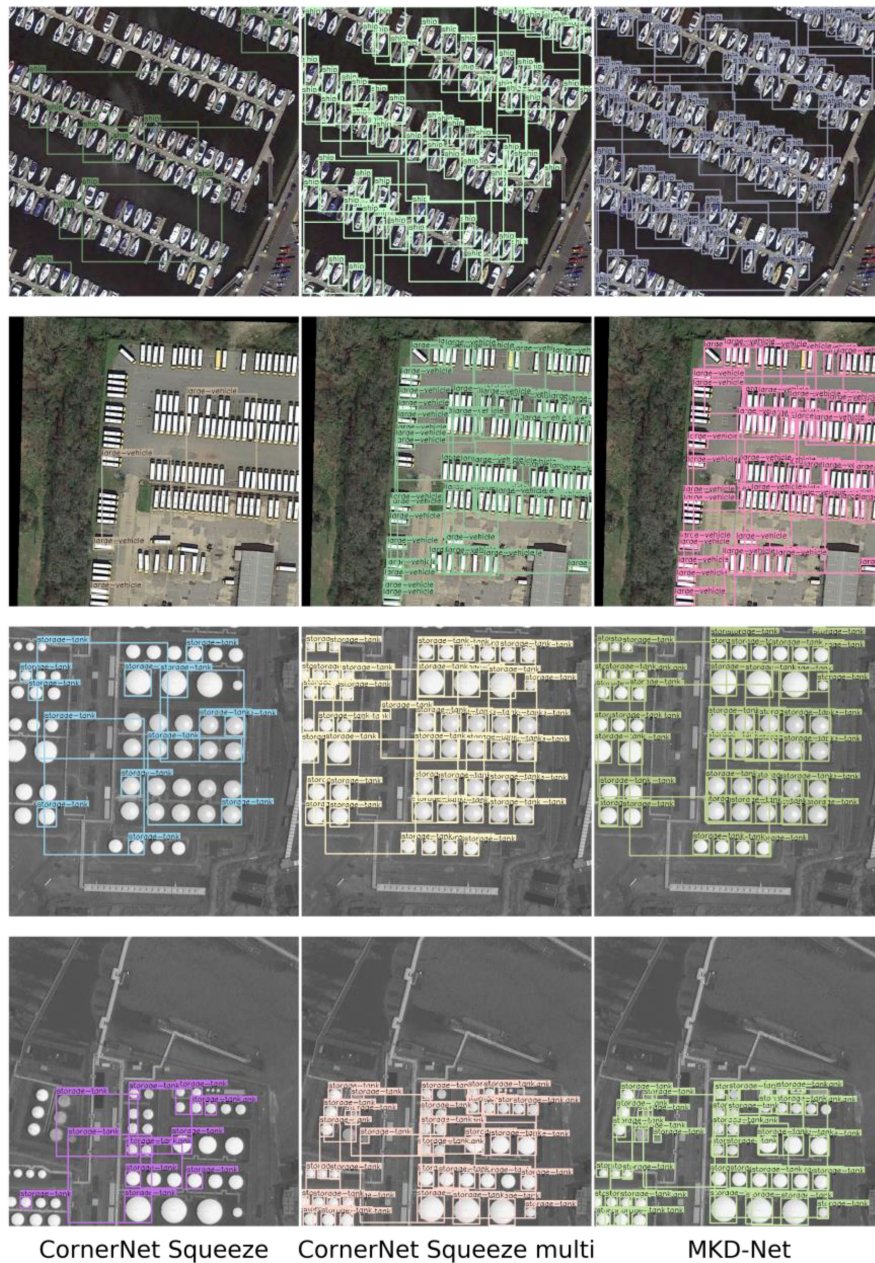


Fig. 7. Example results of the three detectors for dense small objects in images.

of the three models decreases sequentially, and the inference speed increases sequentially. All of them need to predict the heat map for generating corners. Our modification mainly focus on optimize the heat map. Therefore, we could surmise that our modification is also suitable for CornerNet and CenterNet. Both of them could achieve a higher mAP on VOC dataset.

E. Evaluation of the Different Categories

In general, large objects are easier to identify than small objects. CornerNet-Squeeze often fails to detect some very large objects. As shown in Fig. 6, there is only a large soccer ball field in the first row and two large planes in the second row.

The backgrounds of these two images are not complicated, and the outlines of these objects are clear. Intuitively, those objects should be easy to detect. However, both CornerNet-Squeeze and CornerNet-squeeze multifailed to detect them. MKD-Net is effective in detecting such target objects, as it benefits from the extra feature map with larger receptive fields for large objects. The MKD-Net also benefits from the channel attention network for reducing the channel noise in the single-category situation.

The aerial images often contain small, dense objects in some regions. In Fig. 7, each image has more than 100 ground-truth boxes in the same category. During inference, there are plenty of corners of the same category in an image. It is difficult to

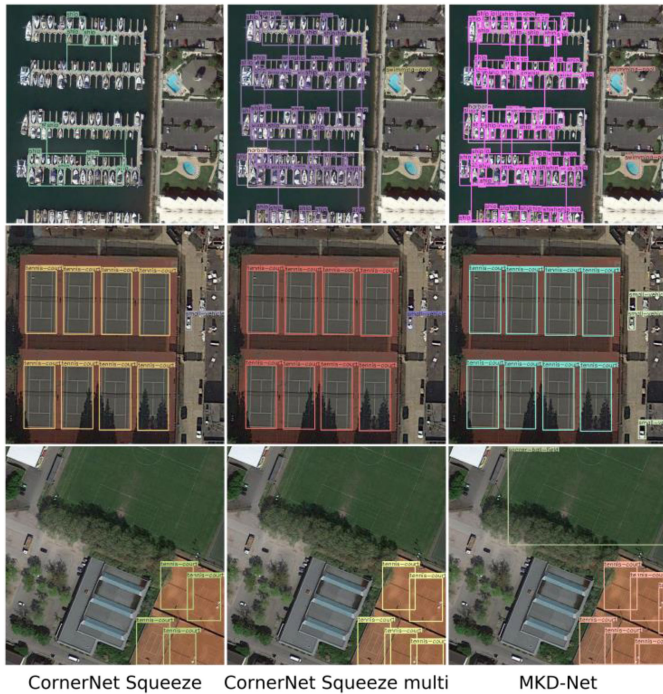


Fig 8. Example results of the three detectors for multiple categories in images.

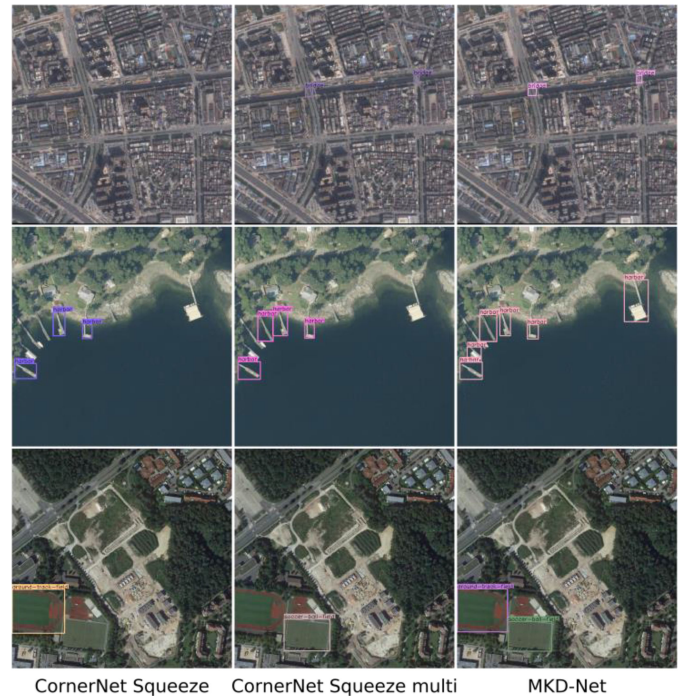


Fig 9. Typical detection results of the three detectors.

group the irrelevant corners into an object. As a result, the embedding module generates numerous paired keypoint errors. Furthermore, CornerNet-squeeze computes the final feature map with an $8\times$ -reduced resolution compared to the original image. The small objects have been subsampled to several pixels in the last layer, leading to difficulty in distinguishing the paired corners of an object. Due to the above two reasons, CornerNet-squeeze missed a large proportion of ground-truth boxes. We can see that CornerNet-squeeze multi and MKD-Net perform better than CornerNet-squeeze. These two methods generate bounding boxes from each feature map and select the top k boxes separately before performing the NMS operation. It would be helpful for the model to capture the information learned at each feature map.

Fig. 8 shows a particular type of scene. Each image has two categories of objects. One of these two categories occupies a primary place with large quantities, such as the ship and tennis court in the figure. Apart from the dominant category, they have some small objects, such as swimming pools and small vehicles. In this scene, a few small objects are detected by CornerNet-squeeze and CornerNet-squeeze multi. Both of them miss the two swimming pools in the first row. Only one small vehicle in the second row is detected by those two methods.

Finally, we provide some visualization results in Fig. 9 to show that MKD-Net indeed enjoys a strong ability to improve the precision of detection. In the first row, both MKD-Net and CornerNet-squeeze multi are able to accurately locate the positions of bridges in a complex background. The other two methods miss part of the harbor in the second row. Each method misses one object in the third row. Only MKD-Net detects all the ground-truth boxes in the last two rows.

V. CONCLUSION

In this article, we have presented an end-to-end multiscale keypoint-based detector for objects in aerial images. Considering the size variability and dense small objects, a fusion module with multiscale features was added. The module fused features from different layers and generated two different feature map resolutions. Both feature maps were used to generate bounding boxes as a pair of corners. Moreover, we proposed a supervised category attention network to predict the probability of a category of objects being contained in an image. The output of the network was used as class weights to adjust the corner distribution heat map value. In the case of a small number of object categories in a test image, this module can weaken the influence of channel noise in feature maps and reduce misclassified objects. The experimental results on the DOTA and PASCAL VOC dataset demonstrate the competitive results of the method.

REFERENCES

- [1] S. Ren *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, pp. 91–99, 2015.
- [2] T. Y. Lin *et al.*, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2017.
- [3] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," 2018, *arXiv:1808.01244*.
- [4] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [5] H. Law *et al.*, "CornerNet-lite: Efficient keypoint based object detection," 2019, *arXiv:1904.08900*.
- [6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1440–1448, 2015.
- [7] T. Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, pp. 936–944, 2017.

- [8] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, pp. 21–37, 2016.
- [9] K. Duan *et al.*, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 6568–6577, 2019.
- [10] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 840–849, 2019.
- [11] J. Wang *et al.*, "Region proposal by guided anchoring," *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2960–2969, 2019.
- [12] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9626–9635, 2019.
- [13] T. Kong *et al.*, "FoveaBox: Beyond anchor-based object detector," *IEEE Trans. Image Process.*, pp. 7389–7398, 2020.
- [14] G. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 3974–3983, 2018.
- [15] P. Zhu *et al.*, "Vision meets drones: A challenge," 2018, *arXiv1804.07437*.
- [16] G. Cheng *et al.*, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014.
- [17] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, Aug. 2017 pp. 3652–3664.
- [18] L. W. Sommer, T. Schuchert, and J. Beyerer, "Fast deep vehicle detection in aerial images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pp. 311–319, 2017.
- [19] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–864, Jun 2019.
- [20] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [21] X. Yang *et al.*, "SCRDet: Towards more ROBUST detection for small, cluttered and rotated objects," *ICCV*, pp. 8231–8240, 2019.
- [22] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving Object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5561–5569.
- [23] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2274–2284.
- [24] A. Newell and J. Deng, "Pixels to graphs by associative embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2168–2177.
- [25] K. Li, G. Cheng, and S. Bu, "Rotation-Insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2017.
- [26] G. Cheng, P. Zhou, and J. Han, "RIFD-CNN: Rotation-invariant and fisher discriminative convolutional neural networks for object detection," in *Proc. Comput. Vis. Pattern Recognit.*, pp. 2884–2893, 2016.
- [27] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
- [28] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [29] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.



Jinhe Su received the Ph.D. degree in computer science from University of Chinese Academy of Sciences, Beijing, China, in 2018.

He is currently a Lecturer with Computer Engineering College, Jimei University, Xiamen, China. His research interests include deep learning, object detection/recognition, and image/video retrieval.



Jiajia Liao is working toward the graduate degree in multimedia information processing at Computer Engineering College, Jimei University, Xiamen, Fujian, China.

Her research interests include high-resolution remote sensing image object detection, image processing, machine learning, deep learning, and computer vision.



Dujuan Gu received the Ph.D. degree in computer science from University of Chinese Academy of Sciences, Beijing, China, in 2015.

She has been a Postdoctoral Fellow with Tsinghua University, Beijing, China. She is currently a Research Member at NSFOCUS. She is an IEEE member and a CCF member. Her current research interests include cloud computing, network security, computer architecture, and data mining.



Zongyue Wang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2012.

He is currently an Associate Professor with Computer Engineering College, Jimei University, Xiamen, China. His research interests include point cloud processing, machine learning, object detection/recognition, and image/ video retrieval.



Guorong Cai received the Ph.D. degree in artificial intelligence from Xiamen University, Fujian, China, in 2013.

He is currently an Associate Professor with Computer Engineering College, Jimei University, Xiamen, China. His research interests include 3-D reconstruction, point cloud processing, machine learning, object detection/recognition, and image/video retrieval.