How to Fully Exploit The Abilities of Aerial Image Detectors

Junyi Zhang
Sun Yat-Sen University
zhangjy329@mail2.sysu.edu.cn

Junying Huang
Sun Yat-Sen University
huangjy229@mail2.sysu.edu.cn

Xuankun Chen
Sun Yat-Sen University
chenxk3@mail2.sysu.edu.cn

Dongyu Zhang*
Sun Yat-Sen University
zhangdy27@mail.sysu.edu.cn

Abstract

Detecting objects in aerial images usually faces two major challenges: (1) detecting difficult targets (e.g., small objects, objects that are interfered by the background, or various orientation of the objects, etc.); (2) the imbalance problem inherent in object detection (e.g., imbalanced quantity in different categories, imbalanced sampling method, or imbalanced loss between classification and localization, etc.). Due to these challenges, detectors are often unable to perform the most effective training and testing. In this paper, we propose a simple but effective framework to address these concerns. First, we propose an adaptive cropping method based on a Difficult Region Estimation Network (DREN) to enhance the detection of the difficult targets, which allows the detector to fully exploit its performance during the testing phase. Second, we use the well-trained DREN to generate more diverse and representative training images, which is effective in enhancing the training set. Besides, in order to alleviate the impact of imbalance during training, we add a balance module in which the IoU balanced sampling method and balanced L1 loss are adopted. Finally, we evaluate our method on two aerial image datasets. Without bells and whistles, our framework achieves 8.0 points and 3.3 points higher Average Precision (AP) than the corresponding baselines on VisDrone and UAVDT, respectively.

1. Introduction

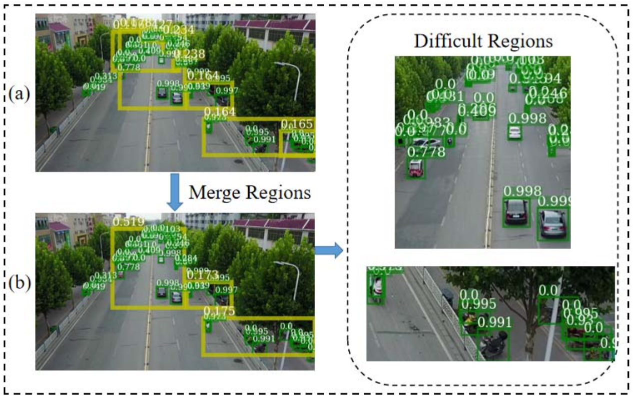

Figure 1: Find difficult regions on the training set. We use sliding window search to obtain regions and calculate their scores based on the prediction results from a preliminary trained detector, as shown in in (a). Then, we merge the regions, the merged regions are displayed in (b). Finally, we select the top N regions with the highest score as the difficult regions, which can be used to train the difficult region estimation network.

Object detection in aerial images has attracted significant attention worldwide due to its important application potential in traffic monitoring [21, 32] and pedestrian tracking [28]. In previous years, the works are mostly based on the sliding window search [27, 1] and the handcrafted features [21]. Recent years, with deep learning becoming the dominative technique in object detection, the related detectors (e.g., R-CNN series [26, 11, 9, 8], YOLO series [23, 24, 25], SSD series [19, 7], etc.) have achieved huge success in image detection of natural scenes. Different from the natural images (e.g., images in Pascal VOC [6] and MS COCO [17]), the aerial image has several unique characteristics: (1) the objects are generally small; (2) uneven distribution of objects; and (3) the angle of camera shooting is not fixed. Due to the influence of these characteristics, the detectors designed for natural images often encounter many

*Corresponding author is Dongyu Zhang. This work was supported in part by Natural Science Foundation of China under Grant No. 61876224, in part by the Natural Science Foundation of Guangdong Province under Grant No. 2017B010116001, and in part by the Fundamental Research Funds for the Central Universities.

difficult targets in the detection of aerial images, this is the first challenge. Another challenge is the inherent imbalance problem in object detection, which hinder the model from optimizing in the best direction during training, this adverse reaction is especially noticeable on unbalanced aerial images. Due to the existence of the above two major challenges, the detectors which perform well in natural images often perform poorly in aerial images in terms of speed and accuracy.

In order to compensate for the weakness of the general detector, many detectors specifically designed for the detection of the aerial image have been proposed. For example, some detectors [12, 5] aim to improve the detection performance on small objects. They usually enhance the effective representation of small objects by carefully designing and transforming deep network structures or using feature fusion of different scales [14, 20]. However, in the case where the image is large-resolution and with large-density of small objects, the results obtained by these methods are mostly unsatisfactory in accuracy. Besides, some researchers propose to make improvements on the anchor, such as designing rotated anchors [5] or proposing the guided anchoring scheme [31]. But, these detectors still do not truly solve the thorny problem of detecting difficult targets, because they only improve the detection of a kind of targets. In general, most of the above detectors are dedicated to improving the structure of the model to achieve better results, which is limited for the improvement of the effect, because the model does not fully exert their abilities during training and testing. Therefore, for the existing feature extraction models and classifiers, we consider that the improvements can be made during the training and testing phases to fully exploit the abilities of aerial image detectors.

Attempting to address these concerns, we propose a simple but effective framework. For the testing phase, just like a teacher teaches students, the teacher should pay more attention to students with poor grades to improve the overall performance of all students. Inspired by this motivation, we propose a *difficult region estimation network (DREN)* to estimate the difficult regions and then retest these difficult regions. We call the regions where the difficult-to-detect objects are concentrated as difficult regions. For the training phase, in order to alleviate the impact of the imbalance problem, we adopt the IoU balanced sampling method [22] and the balanced L1 loss [22] in balance module. Furthermore, we use the trained DREN to generate some effective training data, such data enhancement is important for training a powerful detection model. In general, the main contributions of this work are as follows:

- We provide a simple but effective framework which can fully exploit the abilities of the aerial image detectors by enhancing the detection of difficult targets and alleviating the impact of the imbalance problem.
- Extensive experiments and evaluations on two aerial image datasets demonstrate the validity and stability of our framework.

2. Related Work

Natural Image Detection. Object detection is an active research topic in the computer vision field. Generally speaking, object detection refers to the detection of natural images, also known as general object detection. The existing general object detection methods can be divided one-stage and two-stage. The one-stage detectors includes SSD [19], YOLO [23], and RetinaNet [16]. The two-stage detectors includes Fast-RCNN [8], Faster-RCNN [26], and Mask-RCNN [11]. We mainly introduce R-CNN [9] and a sequence of later works which are developed based on it. R-CNN [9] is the first work of the R-CNN series, it adopts the selective search algorithm [30] to get the candidate boxes and use SVM as classifier. Fast R-CNN [8] accelerates R-CNN by introducing an ROI pooling layer. Faster R-CNN [26] further improves the speed and accuracy by introducing a learnable network to replace the proposal generation stage. Later, Mask R-CNN [11] achieves the state-of-the-art performance by adding a segmentation branch. Recently, anchor free methods are also very popular for the detection of natural image. Cornernet [13] is an anchor free work in object detection, it detects an object bounding box as a pair of key points. [4] is an improved anchor free work based on Cornernet later. However, natural images are very different from aerial images, so these detectors cannot be used directly for aerial image detection.

Aerial Image Detection. Different from natural images, aerial images have several unique characteristics. Therefore, a lot of detectors specifically for aerial image detection have been proposed for a long time. Here, we only introduce some methods based on deep learning, because they are more related to our work. In [18], a fast multi-class vehicle detection approach on aerial images is proposed. [12, 5, 14, 20] aim to improve the detection performance on small targets in aerial images by carefully designing and transforming deep network structures or using feature fusion of different scale. Besides, some approaches make improvements on the anchor, such as designing rotated anchors [5] and proposing the guided anchoring scheme [31]. In [33], a framework for clustered region detection has been proposed, we are inspired by the work. Different from ClusDet [33], our method is to consider the regions where the difficult targets are concentrated, and we abandoned ScaleNet of ClusDet to streamline the entire process. Besides, numerous algorithms for aerial image detection are discussed in [34].

Imbalance in Object Detection. In addition to an excellent structure, a detector also needs to be well trained to perform at their best. However, the imbalance will pre-

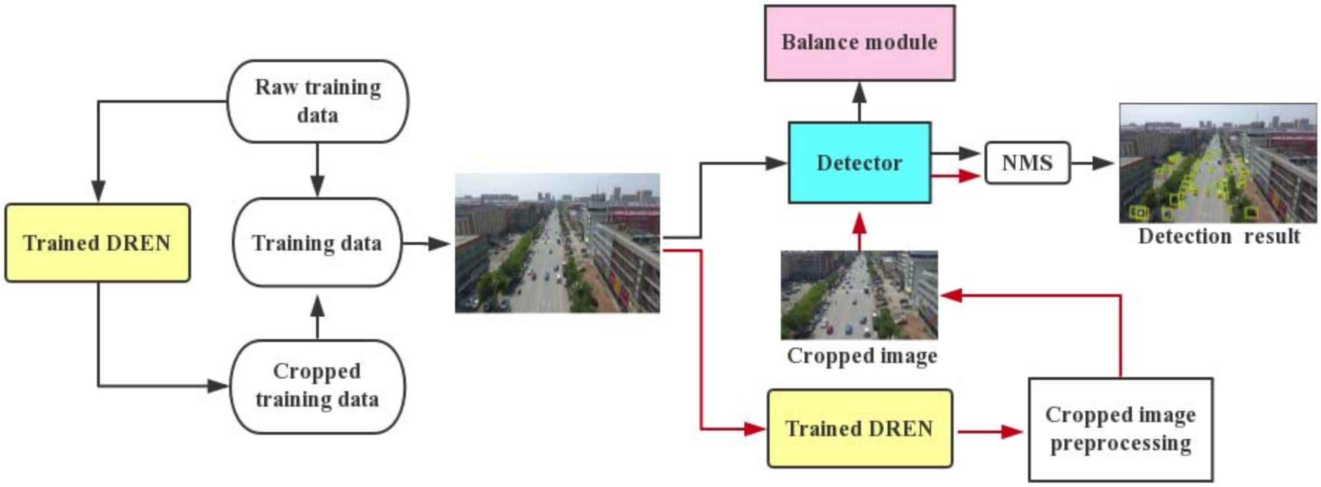

Figure 2: This is our framework. The trained Difficult Region Estimation Network (DREN) is used to estimate the difficult regions during testing and generate some cropped training images during training. When testing, both the original image and the cropped images are passed to the detector, then the generated candidate boxes are merged together for NMS. The balance module contains IoU-balanced sampling and Balanced L1 Loss, which is only used during training and the details of it are shown in Figure 3.

vent the power of well-designed model architecture from being fully exploited. For example, at the sample level, the hard samples are more valuable for training but are usually less, so this should be taken into account when sampling. OHEM [29] is a popular hard mining method, which can help drive the focus towards hard samples, but it is sensitive to noise labels. Focal loss [16] also can alleviate this problem in one-stage detectors, but is found little improvement used in two-stage detectors. Libra R-CNN [22] is a recent proposed framework towards balanced learning for object detection, which integrates IoU-balanced sampling, balanced feature pyramid, and balanced L1 loss. We adopt IoU-balanced sampling and balanced L1 loss from Libra R-CNN in our proposed framework to alleviate the impact of imbalance without introducing additional testing time.

3. Proposed Method

3.1. Overview

As shown in Figure 2, our proposed framework mainly contains a DREN and a balance module. When testing, both the original image and the cropped images are passed to the detector, then the generated candidate boxes are merged together for NMS. When training, we adopt IoU-balanced sampling method, the balanced L1 loss, and use the trained DREN to generate more diverse training data to alleviate imbalances at different levels of our framework.

3.2. Difficult Region Estimation

We train a preliminary detector in advance to get the predicted boxes and scores on the training set. Then we use the predicted box and score to calculate the score for the regions. Finally, we merge the regions with intersections to get the final difficult regions and use these difficult regions to train the DREN.

Calculate Scores of Regions. We use sliding window search to obtain regions. The formula for calculating the score for each region is as follows:

$$\begin{aligned} M &= \frac{\sum\limits_{i \in N} score_i}{N} \\ S &= \frac{N^{\frac{3}{2}} \times \sqrt{M}}{A} \end{aligned} \tag{1}$$

where p is a region, N is the number of the predicted boxes in p, M is the average of the scores of all the boxes in p, A is the area of p, S is the final score of p. Based on experimental experience, we set the ratio (N/M) as 3:1 to balance the magnitude of N and M.

Merge Regions. In order to dig out a continuous region of difficult target aggregation, after calculating the scores of all regions, we merge some regions with intersections to get the final difficult regions. The specific process of merging is shown in Algorithm 1.

3.3. Alleviate Imbalance

Generate Cropped Images. Since the aerial image is of large resolution, or the density of the objects is large, or the

Algorithm 1 IoU-Based Iterative merge (IIM)

```
Input: the set of regions $S$, merge-threshold $t$, maximum
    number of merged regions $N_{max}$
Output: merged regions $S'$
function IIM($S,\ t,\ N_{max}$)
    $S' \leftarrow S$
    while $|S'| > N_{max}$ do
        $mious \leftarrow ComputeIoUs(S', MaxIous)$
        if $max(mious) < t$ then
            break
        else
            $x,\ y \leftarrow argmax(mious)$
            $S' \leftarrow S' - x - y + merge(x,\ y)$
        end if
    end while
    return $S'$

function MaxIoUs($box1, box2$)
    $area1,\ area2 \leftarrow Area(box1),\ Area(box2)$
    $area0 \leftarrow Area(box1 \cap box2)$
    return $area0 / \min(area1, area2)$
```

distribution of the objects is uneven, using random cropping method to enhance training data is not appropriate. Therefore, we propose to generate more representative training images using the trained DREN. In our experiments, four cropped images (the top four difficult regions in score) are generated for each image, the entire training dataset is four times larger than the original dataset.

IoU-balanced Sampling. As [22] mentioned, there are more than 60% hard negatives have an overlap greater than 0.05, but random sampling only provides us 30% training samples that are greater than the same threshold. IoU-balanced sampling is a piecewise sampling method which can resolve the above contradiction. Suppose we need to sample N negative samples, then we evenly split the sampling interval into K parts according to IoU. The number of selected samples for each part is

$$P = \frac{N}{K} \tag{2}$$

where K is 3 in our experiments, the range of IoU is [0,0.3].

Balanced L1 Loss. In addition to IoU-balanced sampling, we also adopt balanced L1 loss [22] in the balance module, which is denoted as L_b. Balanced L1 loss is derived from the conventional smooth L1 loss, the large gradients are clipped with a maximum value of 1.0. The promoted gradient formulation is designed as:

$$\frac{\partial L_b}{\partial x} = \begin{cases} x = & \alpha \ln(b|x|+1) \quad if|x| < 1 \\ y = & \gamma \quad otherwise, \end{cases} \tag{3}$$

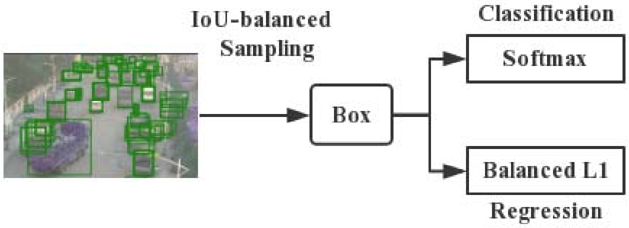

Figure 3: The details of the balance module in Figure 2. This is the process of sampling the candidate box and calculating the loss during training.

By integrating formulation (3), the balanced L1 loss can be expressed as:

$$L_b(x) = \begin{cases} \frac{\alpha}{b}(b|x|+1)\ln(b|x|+1) - \alpha|x| & if|x| < 1 \\ \lambda|x| + C & otherwise, \end{cases} \tag{4}$$

where the parameters λ, α, and b are constrained by

$$\alpha \ln(x+1) = \lambda \tag{5}$$

We set α = 0.5 and λ = 1.5 in our experiments.

3.4. Cropped Image Preprocessing.

According to the experience in our experiments, the aspect ratio of the cropped image can not be too large or too small. So we do the following preprocessing on the cropped image: if the width (height) of the cropped image is greater than 0.7×$W(H)$, we cut the cropped image into two equal parts, then extend the images according to the extension rules. If the width (height) of the cropped image is smaller than 0.6×$W(H)$, we expand it to 0.6×$W(H)$. W and H represents the width and height of the original image. In the experiment, we found that retaining some information around the target is helpful to detect the target, which is why the image is expanded.

4. Experiments and Results

4.1. Implementation Details

Our detector are implemented on PyTorch and Detectron [10]. The Mask R-CNN [11] with Feature Pyramid Network (FPN) [15] are adopted as the baseline detection network. We train our detector for 90k iterations on 8 TITAN Xp GPUs. The initial learning rate is 0.001, after 60k iterations, the learning rate decreases to 0.0001. A momentum of 0.9 and parameter decay of 0.00001 are used. The difficult region estimation network (DREN) is implemented on network of SSD [19] and parameters are set according to it. The other parameters of our detector not specified are in accordance with the initial settings of the Detectron. When

Table 1: The detection results on the validation set of VisDrone.

Base	Method	pedestrian	people	bicycle	car	van	truck	tricycle	awning-tricycle	bus	motor	**AP**
ResNeXt 152	Baseline	17.7	10.2	9.3	49.8	28.7	24.8	17.9	10.4	36.2	18.1	22.3
	Baseline+UC (2 × 2)	19.8	10.1	8.6	51.1	25.6	19.4	17.5	8.9	34.4	18.2	21.3
	Baseline+RC (4)	22.4	13.4	10.7	51.4	28.1	25.5	20.4	10.4	37.5	22.3	24.2
	Baseline+Ours	**25.7**	**17.6**	**16.8**	**55.6**	**38.0**	**34.5**	**24.6**	**14.6**	**49.8**	**25.8**	**30.3**
ResNeXt 101	Baseline	14.3	8.1	8.1	46.8	26.9	22.7	14.6	7.0	34.8	15.0	19.8
	Baseline+UC (2 × 2)	17.2	8.1	6.3	46.2	26.0	18.8	14.0	6.5	33.7	15.2	19.2
	Baseline+RC (4)	17.8	10.7	8.8	46.3	27.1	23.4	15.7	7.7	35.0	17.8	21.0
	Baseline+Ours	**22.2**	**15.0**	**14.5**	**53.8**	**34.5**	**30.7**	**21.0**	**11.2**	**45.2**	**22.7**	**27.1**

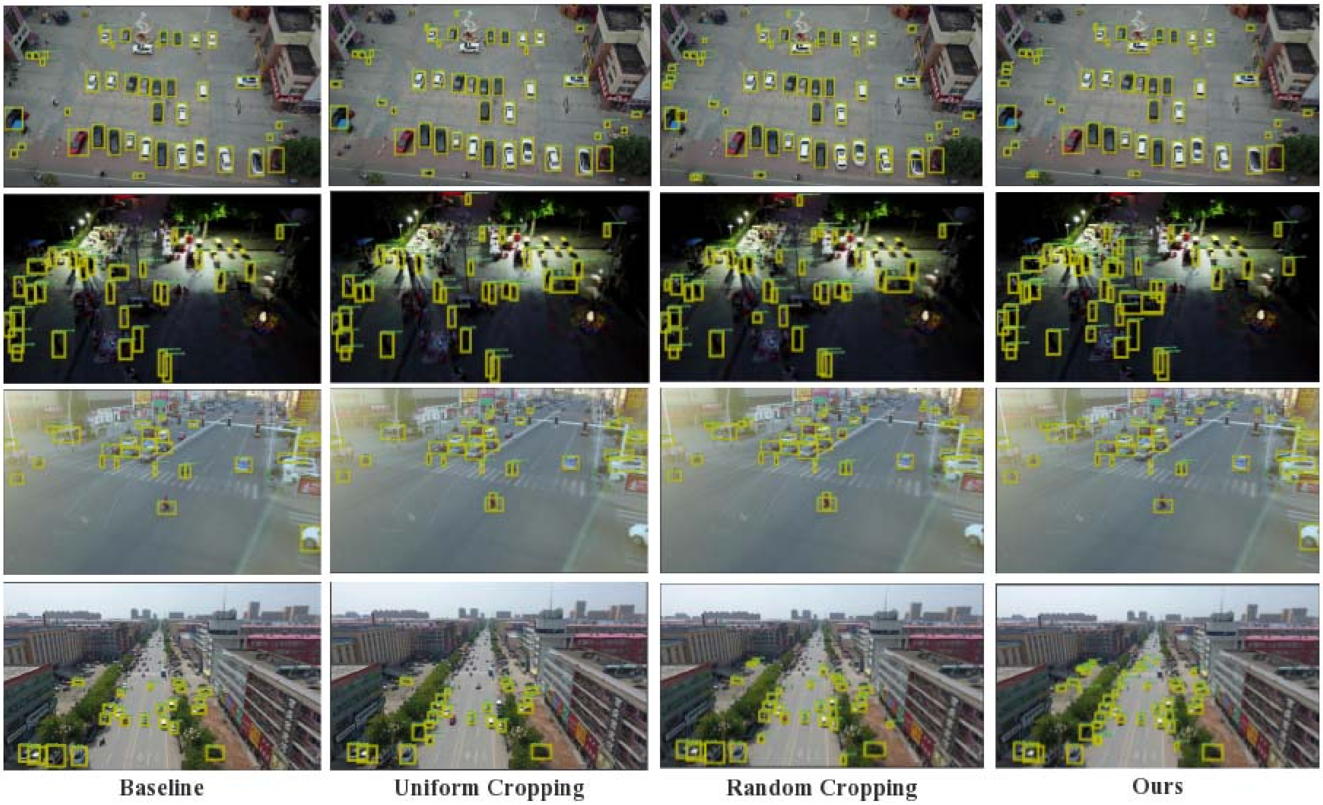

Figure 4: This is the visualization of the detection results on the validation set of VisDrone. For each test image, we show the visualized result of baseline, uniform cropping, random cropping, and our proposed method.

testing, the candidate boxes predicted in the cropped image will be deleted if they are on the edge. The number of difficult regions generated by DREN for each original test image is 3 by default in the experiment.

4.2. Datasets and Evaluation Metric

We evaluate our approach on VisDrone [35, 34] and UAVDT [3]. Next, we briefly introduce the datasets and the evaluation metric below:

Evaluation Metric. We follow the evaluation protocol in MS COCO [17]. We use the Average Precision (AP) metric to evaluate the detection results. The AP is averaged over multiple Intersection over Union (IoU) values. Specifically, we use ten IoU thresholds of [0.50:0.05:0.95].

VisDrone. This is also an aerial image dataset which consists of 6471 images in the training set, 548 images in the verification set and 3190 in the test set. The resolution is about 2000×1500 pixels. The images in the training set and verification set have rich annotations on ten categories of objects. In this dataset, the density of objects is large and the objects are unevenly distributed. Since the organizer of this dataset does not provide the labels for the test set, we use the verification set to test the trained model.

UAVDT. This is an aerial image dataset. It contains approximately 40k representative images including 23258 images for training and 15069 images for testing. The resolution is about 1024×540 pixels. There are three categories of annotation objects, including cars, buses, and trucks.

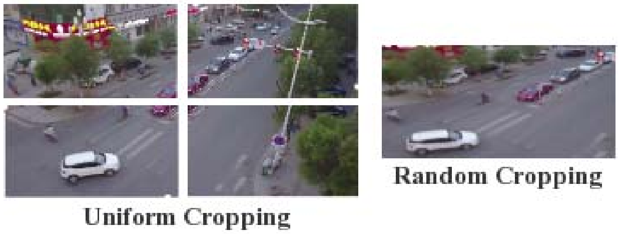

Figure 5: Uniform cropping and random cropping.

4.3. Quantitative Results

We adopt Mask R-CNN [11] with Feature Pyramid Network (FPN) [15] as the baseline model. In addition to the baseline, we also compare our method with uniform cropping (UC) and random cropping (RC). As shown in Figure 5, UC means that the image is evenly divided into four parts. For a fair comparison, the width and height of each randomly cropped image are larger than $0.6 \times W$ and $0.6 \times H$, and the number of randomly cropped images is 4. W and H are the width and height of the original image. Because without the scale and size constraints, the effect of random cropping is very pool.

Table 2: Ablation study of detection result on validation set of VisDrone.

Method	AP
Baseline	22.3
Ours w/o generate training data	28.5
Ours w/o balance module	24.5
Ours w/o cropped image preprocessing	28.7
Ours w/o adaptive cropping	25
Ours (ResNeXt 152, N=3)	**30.3**

VisDrone. Table 1 shows the comparison results of our approach with the methods of baseline, uniform cropping, and random cropping. Our method achieves an AP of 30.3 when the backbone network is ResNeXt 152, achieving 8.0 points improvement over the baseline. Our method significantly improves all individual indicators compared to the other methods. Figure 4 shows the comparison of visualization results. The second, third, and fourth rows indicate that our method has significant improvements in detecting the small objects farther away. The effect of the first row is not obvious, because the angle of this captured image is just below.

Table 3: Parameter analysis of N. N is the number of difficult regions for each test image, the backbone network is ResNeXt 152. When N is large, the test takes a lot of time, so we only analyze the case where N is 2 to 5.

N	2	3	4	5
VisDrone	29.8	30.3	30.2	**30.4**
UAVDT	17.5	**17.7**	17.6	17.5

UAVDT. Table 4 shows the comparison results of our approach with the methods of baseline, uniform cropping, random cropping, and ClusDet [33]. The detection results of Faster-RCNN [26], R-FCN [2], and SSD [19] are obtained directly from ClusDet. From Table 4 we can see that our method achieves an AP of 17.7 when the backbone network is ResNet 101, achieving 3.8 points improvement over the baseline. When the backbone network is ResNet 50, our method achieves an AP of 15.1, achieving 1.4 points improvement over ClusDet and achieving 3.3 points improvement over the corresponding baseline, which shows that our approach achieves the state-of-the-art on this aerial image dataset. From the comparison results we can see that both Baseline + UC and Baseline + RC also can improve performance, but our approach is the most obvious improvement.

Table 4: The detection results on the test set of UAVDT.

Base	Method	car	truck	bus	**AP**
VGG	Faster-RCNN [26]	-	-	-	5.8
ResNet 50	R-FCN [2]	-	-	-	7.0
N/A	SSD[19]	-	-	-	9.3
ResNet 50	ClusDet[33]	-	-	-	13.7
ResNet 50	Baseline	19.9	4.3	11.1	11.8
	Baseline+UC (2×2)	25.2	3.8	8.8	12.6
	Baseline+RC (4)	**25.4**	4.5	10.4	13.6
	Baseline+Ours	24.6	**6.2**	**14.6**	**15.1**
ResNet 101	Baseline	24.8	4.9	12.0	13.9
	Baseline+UC (2×2)	24.9	4.7	15.3	14.9
	Baseline+RC (4)	26.3	**5.8**	14.6	15.6
	Baseline+Ours	**29.2**	5.1	**18.7**	**17.7**

4.4. Analysis

Parameter Analysis. In the difficult region estimation network, we need to set the parameter N to determine the number of difficult regions to be selected for each test image. It is instructive to understand the effect of N on the performance of our framework. We considered four cases when N is 2 to 5. Table 3 shows that when N is 5, our method can get the best performance on the Visdrone dataset, but this does not mean that 5 is the best because the time consuming is positively correlated with N. When N is 3, its performance is almost the same as when N is 5. It can also be seen from the table that our method can get the best performance on the UAVDA when N is 3 or 4.

Ablation Study. In this experiment, we show how each component in our framework affects the final performance. We consider 6 cases: (a) baseline: we adopt Mask R-CNN with Feature Pyramid Network (FPN) as the baseline model; (b) Ours w/o generate training data: do not use the cropped training images generated by difficult region estimation network as training data; (c) Ours w/o balance module: use random sampling instead of the iou-balanced sampling and use L1 loss instead of the balanced L1 loss in the balance module; (d) Ours w/o cropped image preprocessing: remove the cropped image preprocessing part; (e) Ours w/o adaptive cropping: remove the difficult region estimation network; (f) Ours: the full implementation of our method. The AP results are reported in Table 2. It can be observed from the AP results that each component in our framework is of great importance to obtain full improvement in test performance. Among all components, the most obvious impact on the overall performance is the balance module and adaptive cropping, which shows that the combination of the two is very meaningful, and it also proves that our improvement in the training and testing phase is effective for fully exploiting the performance of the model.

4.5. Discussion

The discussion point is why Baseline + UC does not perform as well as baseline on VisDrone. As can be seen from the results in Table 1, in the experiments when the backbone network is ResNeXt 152 or ResNeXt 101, the performance of Baseline + UC will be a little worse. This may be because the density of the objects in the images is very large, such a cropping method easily destroys the shape of the object in the images, which results in reduced the performance. However, the number of objects in the UAVDT is much less than the number of objects in the visdrone, so this method becomes effective in UAVDT. Notably, Baseline + RC works better than Baseline + UC, probably because we constrain the size and aspect ratio of randomly cropped images.

5. Conclusion

In this paper, we presented a simple but effective framework to fully exploit the abilities of aerial image detectors. Our framework addresses two major challenges in the detection of aerial images. During testing, we propose to strengthen the detection of difficult targets using a difficult region estimation network. During training, we alleviate the impact of imbalance by introducing a balancing module. Besides, we use the well-trained DREN to generate more diverse and representative training images for data enhancement. The full experimental results on two aerial image datasets have demonstrated the effectiveness of our method.

References

[1] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, 11(10):1797–1801, 2014.

[2] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[3] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.

[4] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Object detection with keypoint triplets. *arXiv preprint arXiv:1904.08189*, 2019.

[5] K. Duan, D. Du, H. Qi, and Q. Huang. Detecting small objects using a channel-aware deconvolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[7] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[8] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[10] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron, 2018.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[12] R. LaLonde, D. Zhang, and M. Shah. Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4003–4012, 2018.

[13] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[14] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian. Small object detection in unmanned aerial vehicle images using fea-

ture fusion and scaling-based single shot detector with spatial context analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] K. Liu and G. Mattyus. Fast multiclass vehicle detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, 12(9):1938–1942, 2015.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[20] H. Long, Y. Chung, Z. Liu, and S. Bu. Object detection in aerial images using feature fusion deep networks. *IEEE Access*, 7:30980–30990, 2019.

[21] T. Moranduzzo and F. Melgani. Detecting cars in uav images with a catalog-based approach. *IEEE Transactions on Geoscience and remote sensing*, 52(10):6356–6367, 2014.

[22] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2019.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[24] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[25] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[27] W. Shao, W. Yang, G. Liu, and J. Liu. Car detection from high-resolution aerial imagery using multiple features. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 4379–4382. IEEE, 2012.

[28] Q. Shen, L. Jiang, and H. Xiong. Person tracking and frontal face capture with uav. In *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, pages 1412–1416. IEEE, 2018.

[29] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[30] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[31] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin. Region proposal by guided anchoring. *arXiv preprint arXiv:1901.03278*, 2019.

[32] X. Xie, W. Yang, G. Cao, J. Yang, Z. Zhao, S. Chen, Q. Liao, and G. Shi. Real-time vehicle detection from uav imagery. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–5. IEEE, 2018.

[33] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling. Clustered object detection in aerial images. *arXiv preprint arXiv:1904.08008*, 2019.

[34] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu. Vision meets drones: A challenge. *CoRR*, abs/1804.07437, 2018.

[35] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu. Vision meets drones: a challenge. *arXiv preprint arXiv:1804.07437*, 2018.