AMRNet: Chips Augmentation in Areial Images Object Detection

Zhiwei wei, Chenzhen Duan
Harbin Institute of Technology (Shenzhen)
{19S051024,18s151541}@stu.hit.edu.cn

1 Abstrast

Detecting object in aerial image is challenging task due to 1) objects are often small and dense relative to images. 2) object scale varies in a large range. 3) object number in different classes is imbalanced. Current solutions almost adopt cropping method: splitting high resolution images into serials subregions (chips) and detecting on them. However, few works notice that some problems including scale variation, object sparsity exist when directly train network with chips. In this work, Three augmentation methods are introduced. Specifically, we propose a scale adaptive module compatable with all existing cropping method. It dynamically adjust cropping size to balance cover proportion between objects and chips, which narrows object scale variation in training and improves performance without bells and whistels; In addtion, we introduce mosaic effective sloving object sparity and background similarity problems in areial dataset; To balance catgory, we present mask resampling in chips providing higher quality training sample; Our model achieves state-of-the-art perfomance on two popular aerial images datasets of VisDrone and UAVDT. Remarkably, All methods can independent apply to detectiors increasing performance steady without the sacrifice of inference efficiency.

2 Introduction

The object detection in aerial images has widely application, including smart cities assistance, traffic monitoring, disaster search and rescue due to flexible shooting view and wide receptive field. Many effective solutions have been proposed in nature scene detection(e.g., Faster RCNN [1], RetinaNet [2], SSD [3]). However, aerial datasets have special challenges different from nature images in COCO [4] and Pascal VOC [5] datasets. Aerial detectors get bad perfomance when applying the same strategies with nature images.

These characteristics raise severe problems in areial images object detection: (1) Images are general high spatial resolution, and most of objects are small scale relative to the image. (2) Object scale varies in a widely range due

to the change of shooting angle and elevation. (3) Object number in different category is imbalance in many case.

"Cropping and then detection" is proposed to deal with the problems of small object. More specifially, detector first crop high resolution images into several subregions, denoted as chips, and detect on them. The finnal result is fused by chips and original images. For most of cases, cropping strategies help to improve detection accuracy of small object, because compared with the original images, chips get bigger zoom fator when resize to fixed resolution in inference. Small objects are easily detected when chips are rescaled in high resolution.

In [8], the authors split the images uniformly to show power in small object detection. [9] use K-means to generate object gathering clusters and train a network to predict cluster regions. The work of [10] introduce object density map to discrible object distribution and crops connected region. [11] predict potential hard regions and detect on them.

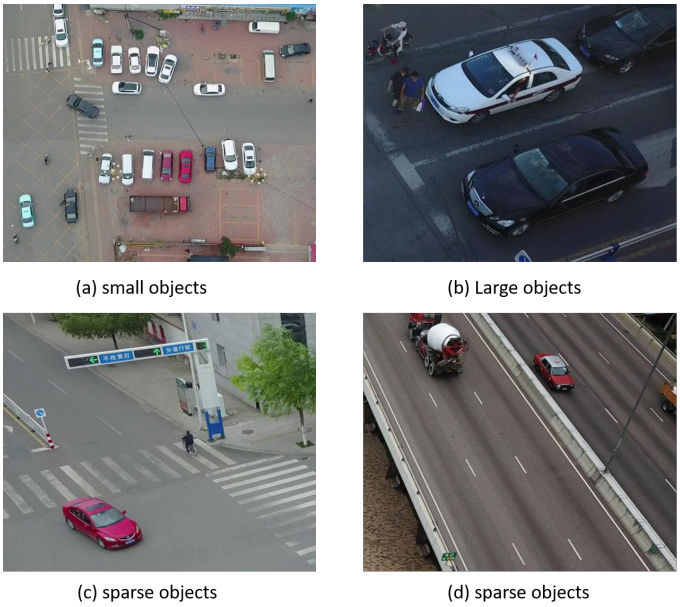

Figure 1: (a)(b) show the scale variation in a large range. (c)(d) present sparse samples which contain few objects

Compared with the original images, object have smaller scale variation inside chips due to the similar magnitude of scale in certain local region. However, severe scale variation still exists among different chips. Training network with those chips which exist scale gap will degrade detector performance [12, 13]. In addtion, some chips are sparse samples which contain few objects due to objects noneuniformly distributed in areial images and shortcoming of cropping method. As show in figure 1, object scale varies from extremely small to relatively large. Some chips contain many background information but less foreground object.

A nature idea to alleviate scale gap is individual resizing chips in training. Assigning a appropriate zoom factor for every chip, makes object scale as similar as possible. We denote the average object scale as the chip scale. However, the current training processes is first resizing samples in a fixed resolution and feed into network. The zoom factor is determined by the size of chips. Compatible with training setting, an feasible method is dynamic adjust chip sizes to affect zoom factor. For low quality sparse sample, intuitive measures is pasting object in chips, increasing object number and balance object number in different class. Motivated by [21], a alterative method is cropping meaining regions from chips and combines them in a new image.

In this paper, we introduce three augmentation methods to relieve problems including scale variation, object sparsity, and class imbalance for areial detectors based on cropping ideas. We propose a scale aware module, called adaptive cropping, which is compatible with existing cropping approach. It dynamically enlarge or narrorw chip size according object average scale. To improve sample quality, we introduce mosaic to augment dataset and effective slove object sparsity and background similarity problems. In addtion, we resample object masks in chips to balance class. We abbreviate our network as AMRNet due to three augmentation methods: adaptive cropping, mosaic augmentation, mask resampling.

We evaluate our approachs on two popular areial images detection datasets: VisDrone and UAVDT. On both datasets AMRNet significantly outperforms prior arts detectors with a large margin. Remarkably, All methods can independent apply to detectiors increasing performance steady without the sacrifice of inference efficiency

In summary, the contributions of our works are:

- A scale adaptive cropping method, which is compatible with all current existing cropping approachs. Without bells and whistles, it improves detector performance directly.
- First introduce mosaic augmentation into aerial images detection. It effectively solves object sparsity and background similarity problems in Visdrone [14] and UAVDT [15] dataset.
- A mask resampling method to relieve class imbalance: pasting mask in images and adjusting catgory, scale, lumination of mask according local context information.
- State-of-the-art object detection performance on VisDrone [14] and UAVDT [15] dataset.

3 Relate work

3.1 General object detection

Detection can be categoried in two class: region-base and region-free. The region-base family, including Faster RCNN [1], Mask RCNN [16], Cascade RCNN [17], first predicts object potenial region and then detects on those proposal. Region-base detectors perform well in object classfication and location but too time consuming. So a series of region-free detectors e.g. SSD [3], YOLO [**?**], RetinaNet [2] appear getting faster detection speed at the cost of degenerative accurateness. Recently some point-base detectors (e.g., cornerNet [18], repPoints [19]) show their special advantage (anchor-free) in object detection. All above mentioned detectors mainly focus on natural images and get bad perfomance when directly apply in aerial images.

3.2 Object detection in aerial images

Comparing with detection in nature images, there are more challenges in aerial images. First, many objects are in small scale relative to high-resolution images. Second, severe scale variation caused by the change of viewpoint and elevation exists in dataset. Third, class imbalance is common in many aerial datasets. Recent works attempt to solve those problems to improve detection performance.

sub-region detection. An effective and leading method to alleviate scale variance and small object detection problem is cropping high resolution images into series chips and detection on them. Many reseachers have detected object on image sub-region and studied how to reasonable cropping image [6, 7, 9, 10, 11]. In[8], the authors uniformly crop images into six chips for detection and show validation in small object detection. The work of [9] crops images by training a netwrok to predict potential object cluster region. The method in [10] introduces object density map and crops connected regions. The approach of [11] trains a network to prdict hard region and detects on it. Above works narrow scale variances in chips compared with the original image, but not consider scale variances among chips. Training network with those "scale gap" chips will degenerate detection performance [12, 13]. Therefore, we propose an adaptive cropping method, which adjusts chips size to balance object scale, which effectively improves detector performance.

Data augmentation. Researchers have implemented many data augmentataion methods including random cropping, ipping, and inputs with multiple scale. In [20], the authors cut a part of image and paste with other image subregion for augmentation. The work of [21] augments dataset by random combining multiple image subregions into a new image. Some special augmentation methods also appear in aerial datasets. The method of [22] crops iamges into four uniform chips to enlarge dataset. The approach of [23] pastes object randomly in images to improve small object detection performance. In [24], the authors take advantage semantic segmentation to paste object on road regions, avoiding

the mismatch of semantic information. Motivate by [21], we introduce mosaic augmentation to aerial images detection and effective slove sparse sample and similar background problem. We adjust object scale into a reasonable range and combine multiple image regions into a masoic image, providing higher quality and quantity training samples.

Class Imbalance. A significant problem in aerial images is imbalance of catgory object number. In [11], the authors use IOU balanced sampling and balanced L1 loss to alleviate class imbalance. The work of [22] divides class into two bags and trains a expert detector seperately. The approach of [24], pastes object several times alleviate class imbalance. We proposal mask resampling method to paste object mask in images. Different from work in [24], we only paste object instance pixel instead of whole ground truth to get more accurate semantic match. In addition, it is unreasonable for different class object pasting in the same scale in [24], so we reconsider strategy of the pasted object scale, illumination and catgory.

4 method

4.1 Overview

Cropping original images into chips is a common method to overcome small obeject and scale varience problems. However, extreme scale gap still exists among chips. Training network with those chips will degenerate performance. In addtion, some chips are sparse samples which only contain few objects. Training network with those samples is low gain and inefficient. The class imbalance also limits the network performance. In this section, we offer some simple augmentation solutions to alleviate above problems and improve detector performance effectively.

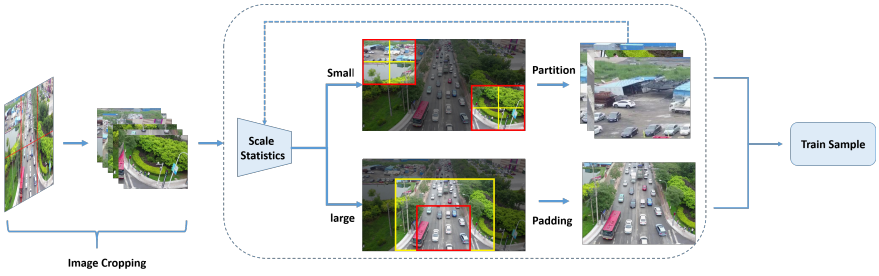

Figure 2: Adaptive Cropping Augmentation. Images first are cropped into uniform chips. For each chip, Object scale information is collected and decides to do partition (top path) or padding (bottom path) operation. The red box and yellow box indicate original chips and adaptive chips separately. The chips from partition iterate processes unless the maximum limit is exceeded or it do padding operation. We collect adaptive chips and reconstruct new training set.

4.2 Adaptive Cropping

An obvious feature of aerial images is a large range of object scale. Due to the change of shooting angle and elevation, objects in same class have 20 times scale difference in VisDrone [14], which is not conducive to network training. Therefore, we propose a scale adatptive cropping method which adjusts cropping size dynamically according object scale.

The processes as figure 2. Images first are croped into chips and then adjust in scale adaption model. We just use simple six uniform cropping as example. We calculated the average object scale for each chips and decide to do a partition or padding operation according scale information. if the object scale is small, the chip will be splited uniformly into four parts. if the scale is large, the chip will be padded. We adjust object area proportion relative to chips, so the object scale is similar when chips resize to fixed resolution in training. We transform object into a narrow and reasonable scale range. Finally, all adjusting chips construct a new training dataset. Following is more rigorous mathematical description.

For a chip, we count object number and object total cover area. Then calculate the average area of the object.

$$avg_{obj} = \frac{area}{num} \tag{1}$$

To transform object in a reasonable scale s(e.g.100), ideal zoom factor φ_{id} get as equation 2.

$$\varphi_{id} = \sqrt{\frac{s^2}{avg_{obj}}} \tag{2}$$

Because the chip resize from $S_w * S_h$ to fix resolution $width * high$ in training. Current zoom factor φ_n is calculated by equation 3.

$$\varphi_n = \min(\frac{S_w}{width}, \frac{S_h}{high}) \tag{3}$$

if $\varphi_n > \varphi_{id}$, it means that the chip require zoom out. The chip new width S_w and high S_h is calculated in equation 4. We enlarge the chip along horizontal and vertical axis evenly, extending reverse at the margin.

$$\begin{aligned} S_w &= \frac{\varphi_{id}}{\varphi_n} * S_w^{'} \\ S_h &= \frac{\varphi_{id}}{\varphi_n} * S_h^{'} \end{aligned} \tag{4}$$

if $\varphi_n < \varphi_{id}$, it means that the chip require zoom in. We split the chip uniformly into four parts. the chip new width S_w and hight S_h is calculated in equation 5.

$$
\begin{aligned}
S_w &= 0.5S_w^{'} \\
S_h &= 0.5S_h^{'}
\end{aligned}
\tag{5}
$$

Exploit above operation iteratively, we can get several adaptive chips which object is in similar scale when train network. In our experiment, we limit the number of maximum partition operation as one, iteration stop after padding operation.

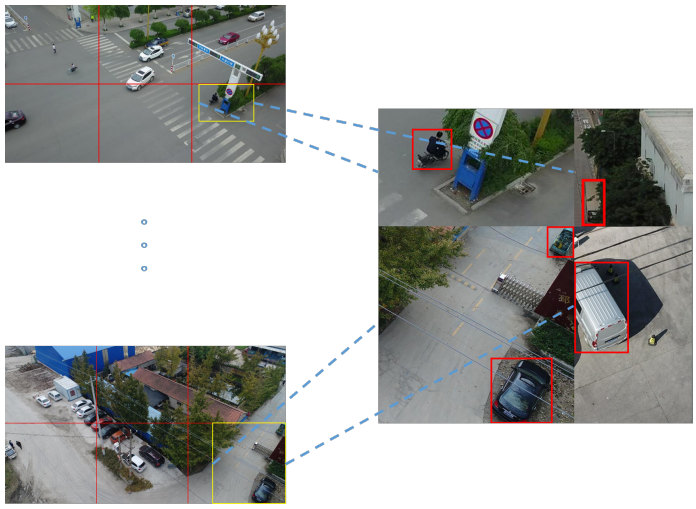

Figure 3: Conbine multiple sparse sample into masoic image. The red line split original image into uniform chips. Some chips are sparse which contains few objects. We crop meaning subregion from chip as yellow box and paste in masoic image after zoom in/out.

4.3 Mosaic augmentation

Some chips are low quality and inefficient to train network. About one fifth chips are sparse samples which contain less three objects when split original images into six parts uniformly in Visdrone [14]. We introduce mosaic augmentation to aerial images detection, combining multiple subregions of sparse samples into a high quality mosaic image. To avoid scale variance problem, we first zoom in/out chips and then use sliding windows to choose appropriate subregion where object is in a reasonable range. The zoom factor directly adopt φ_{id} as equation 2. The processes is shown as figure 3. We crop subregions of sparse samples and paste in masoic image. We also extend the idea to all training sample as data augmentation. We still use sliding windows method and limit the object average scale.

In addation, some chips exist similar background problem and easy overfit. For example, training samples in UAVDT [15] dataset come from series video frame. There are little different between adjacent frame, so images has similar semantic information in trainng samples. Mosaic augmentation is suitable for solving this problem. It combines multiple subregion from different chips and creates complicate context images.

Mosaic images limit object number and scale variance, offering higher quality and quantity training samples. Comparing with original samples, the semantic information in mosaic is more complicate, which forces the network to concentrate on object feature rather than semantic context. It also help detection of objects outside their normal context.

4.4 Mask Resampling

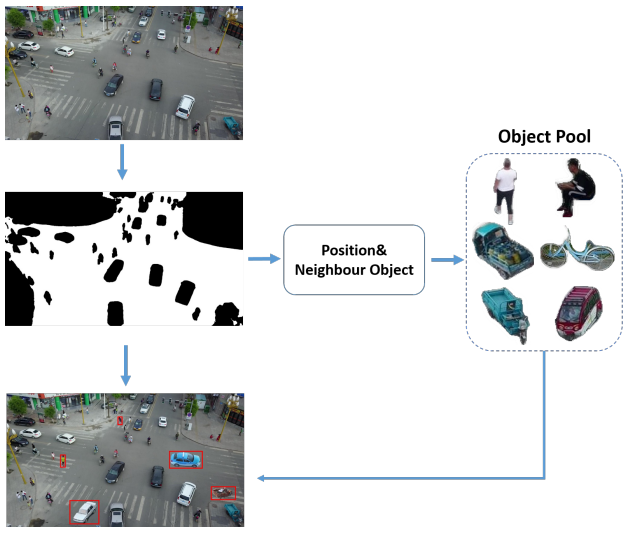

Figure 4: Mask Resampling Augmentation. Image is fed in semantic segmentation network to obtain pastable region. Randomly dertemine pasted position and find nearest object. Choose object in mask pool according neighbours object and paste mask in image.

Another notable problems in aerial datasets is class imbalance. For example, object number of rich catgory over 30 times that of poor catgory in Visdrone [14] and UAVDT [15] datasets. In order to alleviate class imbalance problem, we propose object reasmpling method. We first build a object pool by instance segmentation, and then paste the appropriate objects to training samples. Different from previous work [24] of pasting object ground truth (GT), we only paste object mask and consider catgory, scale, lumination of the pasted object.

We feed aerial images into a COCO pretreated instance segmentation network to get object instance masks. If IOU (Intersection over Union) of the

instance mask and gt is greater than a certain threshold, the gt category is assigned to the mask. We collect these labeled instance mask and build a object pool.

We only paste object in the road regions to ensure semantic correctness. We randomly determine pasted position on the road masks generated by segmentation. Then find the nearest object from pasted position and adjust catgory, scale, lumination of the pasted object. We paste similar catgory with the nearest object. For example, similar catgories of car include van, bus, motor. Because it is reasonable for these objects gather in a local region. The scale of the pasted object P can be calculated by a simple linear function 6 according the nearest object N.

$$\begin{aligned} S_p &= \frac{\overline{S_{pcls}}}{\overline{S_{ncls}}} * S_n \\ \overline{S_{cls}} &= \frac{1}{m} \sum_{i=1}^{n} S_{cls}^i \end{aligned} \tag{6}$$

where S_p, S_n is scale of the pasted and nearest object, $\overline{S_{icls}}$ is the class average scale corresponding class of object i. $\overline{S_{cls}}$ is the class average scale. We also adjust the pasted object lumination close to that of the nearest object in the hsv color space before pasting.

5 experiment

5.1 Implementation details

Our implementation is based on mmdetection tool-boxs [25]. We take uniform cropping as our baseline to validate our approach. Images is uniformly divded into 6 and 4 chips without overlapping on VisDrone [14] and UAVDT [15] datasets. Unless specified, we use retinaNet with feature pyramid network(FPN) as object detector. The input size of detector is 1500 * 800 on the VisDrone [14] and 1000 * 600 on UAVDT [15]. The inference input size of detector is the same with training whenever not specified. We use three scales 1000, 1500, 2000 in multiple scale testing, then fuse all detection boxes with Non-Maximum Suppression(NMS). The fusing threshold is 0.6. Detector is trained for 12 and 5 epochs respectively to two dataset on 2GPUs, each with a bath size of 2. On Visdrone [14] dataset, learning rate sets 0.01 and decreases 0.1 times after the 8th and 11th rounds. On UAVDT [15] dataset, learning rate sets 0.005 and decreases 0.1 times after 4th and 5th rounds.

In adaptive cropping, the expect scale parameter is 100 and 60 in VisDrone [14] and UAVDT [15] with most one partition operation. The reason of setting a specific number is that it is two times average scale of dataset. We remove some oversized crops which have the same length and width as images. In mosaic augmentation, object scale limit over 50 and 30 in Visdrone [14] and UAVDT

[15]. All subregions conbined to 800*800 and 600*600 mosaic image respectively in two datasets. In mask resampling, we paste all catgories except the rich class car.

5.2 Dataset and Evaluation metric

To validate our proposed methods, we evaluate our performance in two publicly datasets: Visdrone [14] and UAVDT [15].

Evaluation Metric. Following the evalutaion protocal on the COCO [4] dataset, we use average precision (AP) as metric. The AP is average precision under different Intersection over Union (IOU) thresholds, ranging from 0.5 to 0.95 with a step size of 0.05. We set 500 for max evelation object in every image.

VisDrone. The dataset consists of 10,209 image in total. More concrete, there are 6471 training images, 548 validation images and 3,190 testing images with 10 categories. The image sclae of dataset is about 2,000 * 1,500 pixels. we evluate performace in validation dataset as existing works beacause of the close of evaluation server.

UAVDT. The dataset contain 23,258 images of training and 15069 iamges in test data with three category including cars,buses and trucks. The resolution of images is about 1080 * 540 pixels. The iamge captured in different scenes and evalation

5.3 Ablation study

To valiate the effectiveness of each modules, we carry out ablation experiments on VisDrone [14] dataset. Resnet50 pretrained in ImageNet [26] is used as backbone to train 12 epoch. The configuration of network is accord with section 4.1. To show contribution clear, we only detect in chips and not fuse bounding box detected in the original image.

Methods	train\test	AP	AP_s	AP_m	AP_l
RetinaNet+FPN	uc	27.0	21.6	35.2	31.7
RetinaNet+FPN+AC	uc	29.5	23.6	39.3	27.6
RetinaNet+FPN	dc	29.3	21.9	40.0	51.5
RetinaNet+FPN+AC	dc	30.5	22.7	41.9	43.6
RetinaNet+FPN*	uc	27.6	22.2	36.0	35.6
RetinaNet+FPN+AC*	uc	31.2	25.9	39.8	36.6
RetinaNet+FPN*	dc	30.6	23.9	40.5	50.8
RetinaNet+FPN+AC*	dc	32.6	25.3	43.5	54.2

Table 1: The ablation on adptive cropping. The 'AC' indicates adaptive cropping augmentation. The 'uc' and 'dc' denotes uniform cropping and density cropping respectively. * represent multple scale testing.

Effect of Adaptive Cropping. We evluate our scale adaptive model based in uniform cropping and density cropping. The experienmental results are

list in table 1. We note that adaptive cropping can boost 2.5 points in uniform cropping, 1.2 points in density cropping. It shows that adaptive augmentation method is robust and easily extend to different cropping method. This promotion is cost free in inference. Moreever, performance improve steeply when use multiple scale testing. Mutiple scale testing increases 0.6 and 1.3 points without adpative model in uniform and density cropping, but increases 1.7 and 2.1 points with adaptive model. The main reason is that adaptive cropping detector focus on detecting object in a certain scale range, and multiple scale testing can help object adjust scale to this interval, so as to give full play to the detector performance.

Methods	AP	AP_s	AP_m	AP_l
RetinaNet+FPN	27.0	21.6	35.2	31.7
RetinaNet+FPN+SR	27.4	22.2	35.6	29.5
RetinaNet+FPN+10k	28.8	23.1	37.2	32.9
RetinaNet+FPN+10k+SR	29.1	23.4	37.7	31.3
RetinaNet+FPN+20k	29.3	23.5	38.0	32.2
RetinaNet+FPN+20k+SR	29.6	24.0	38.4	30.6

Table 2: The ablation on Mosaic augmentation. SR indicates sparse replacement. 10K and 20k is the number of the mosaic augmentation images.

Effect of Mosaic Augmentation. Sparse samples count one fifth in Visdrone when use uniform cropping. We just replace sparse samples to mosaic images and evaluate the performance under the same number of training samples. The experimental result lists as table 2. The performance increases 0.4 points after replacing sparse samples. We extend the idea to all training samples for data augmentation. The sparse replacement still gains steady 0.3 points when add 10K or 20K mosaic images for augmentation. We also study the effect of different augumentation image number. It has 1.8 and 2.3 points increasing when add 10K and 20K mosaic images. The gain gradually decreases as the augmentation images increasing.

AC	MA	MR	AP	AP_s	AP_m	AP_l
			27.0	21.6	35.2	31.7
		✓	28.5	22.4	37.4	32.5
	✓		28.8	48.8	29.4	23.1
	✓	✓	29.0	23.4	37.6	33.1
✓	✓		30.7	24.8	40.6	28.6
✓	✓	✓	30.8	24.6	41.0	29.3

Table 3: The ablation on all models. Images number is 10K in MA.

Effect of Mask Resample. We paste object masks into images for allevating class imbalance. The mask resampling can effectively increase performance

from 27.0 to 28.5. We also study the joint effect for all models. Mask resampling increase less after adopting adative cropping and mosaic augmentation. We deduce mosaic images increase the number of rare class objects, overlapping improvement with mask resampling.

Expect Scale	AP	AP_s	AP_m	AP_l
50	28.4	24.8	33.9	20.0
100	29.5	23.6	39.3	27.6
150	28.2	21.3	38.7	33.8

Table 4: Parameter analysis of Expect Scale in Adaptive Cropping.

The Effect of Hyperparameter Expect Scale In adaptive cropping, we need set the parameter of expect scale to calculate zoom factor. We consider three cases when scale set from 50 to 150 with step 50. It is intuitive that network graduately focus on detecting larger scale objects when set bigger expect scale. AP_s decreases and AP_l increases with scale from 50 to 150 in table 5. We note the best perfomance parameter is 100. The reason we think is scale matching exist between object training and testing. The average scale of VisDrone [14] is 50, and chips are common resized about twice in inference stage. So it is reasonable to set expect scale as twice of average scale.

5.4 Quantitative Result

Mehod	Backbone	Test data	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
ClustDet	ResNet 50	Original+cluster	26.7	50.6	24.7	17.6	38.9	51.4
ClustDet	ResNet 101	Original+cluster	26.7	50.4	25.2	17.2	39.3	54.9
ClustDet	ResNeXt 101	Original+cluster	28.4	53.2	26.4	19.1	40.8	54.4
DMNet	ResNet 50	Original+density	28.2	47.6	28.9	19.9	39.6	55.8
DMNet	ResNet 101	Original+density	28.5	48.1	29.4	20.0	39.7	57.1
DMNet	ResNeXt 101	Original+density	29.4	49.3	30.6	21.6	41	56.9
AMRNet	ResNet 50	Original+UC	31.7	52.7	33.1	23.0	43.4	58.1
AMRNet	ResNet 101	Original+UC	31.7	52,6	33.0	22.9	43.4	59.5
AMRNet	ResNeXt 101	Original+UC	**32.1**	53.0	33.2	23.2	43.9	60.5
ClustDet★	ResNeXt 101	Original+cluster	32.4	56.2	31.6	-	-	-
AMRNet★	ResNeXt 101	Original+UC	**36.1**	60.1	37.0	29.0	45.5	60.9

Table 5: Quantitative result for Visdrone dataset

Visdrone. We evaluate our approach comparing with previous method. The result shows as table 5. To make it fair, we train network under the same configuration with [10] and infer in similar resolution. Our approach achieve state-of-the-art precision. It is noted that we exceed the best accuracy of the previous method with big margin only using resNet50 backbone. We get a

high boost when use multple scale testing. It achieve 36.1 AP and improve 3.7 points compared with ClusDet [9]. The adptive cropping model perform well in multiple scale, so the gain in multiple scale catch up that in single scale.

Mehod	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FRCNN+FPN	ResNet 50	11.0	23.4	8.4	8.1	20.2	26.5
ClusDet	ResNet 50	13.7	26.5	12.5	9.1	25.1	31.2
DMNet	ResNet 50	14.7	24.6	16.3	9.3	26.2	35.2
HFEA	ResNet 50	15.1	-	-	-	-	-
Baseline	ResNet 50	15.2	27.3	15.4	9.4	26.3	36.8
Baseline+Mosaic	ResNet 50	16.8	29.0	17.6	10.7	29.8	31.8
Baseline+OurAll	ResNet 50	**18.2**	30.4	19.8	10.3	31.3	33.5

Table 6: Quantitative result for UAVDT dataset

UAVDT. Images in UAVDT dataset come from a serie video adjacent frames. So images have similar background in trainng sample. We chooes images with step of five and split uniformly in 2*2 crops to reconstruct trainset. We use faster RCNN with FPN training on new dataset as baseline. We add 20K mosaic images to augmentation, offering dataset more complicate semantic images.

The experiment result as table 6. We fuse boxs detected in crops and original images as [9, 10, 11]. It is noted that our baseline achieve high ap compared with previous method. We deduce that the dataset is easily overfitting and it is not nesscessary to training network with all images. Remarkly, we deduece mosaic alleviate background similarity problem and effective boosts 1.6 points compared with baseline. In the end, our methods achieve 18.2 AP with state-of-the-art perfomance.

6 Conclusion

In this paper, we propose three augamenataion methods in aerial images detection: adaptive cropping, masoic augmentation, mask resampling. Adptive cropping alleviates scale variance among chips by adjust the aera propoation of objects to chips. A relatively uniform scale is conducive to network learning. Masoic augmentation sloves object sparsity and background similarity problems, imporving the quality and quantity of training samples. Mask resampling balances the differnt class object number with pasting instance masks. Extend quality result shows our approachs achieves state-of-the-art performance on two popular aerial images detection datasets with large marge. All propose methods are cost free in inference and easily extend to detection based on cropping.

References

[1] 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*, 2015, pp. 91–99.

[2] 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*, 2017, pp. 2980–2988.

[3] 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*. Springer, 2016, pp. 21–37.

[4] 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*. Springer, 2014, pp. 740–755.

[5] 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*, vol. 88, no. 2, pp. 303–338, 2010.

[6] M. Gao, R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Dynamic zoom-in network for fast object detection in large images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6926–6935.

[7] Y. Lu, T. Javidi, and S. Lazebnik, "Adaptive object detection using adjacency and zoom prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2351–2359.

[8] F. Ozge Unel, B. O. Ozkalayci, and C. Cigla, "The power of tiling for small object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[9] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8311–8320.

[10] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 190–191.

[11] J. Zhang, J. Huang, X. Chen, and D. Zhang, "How to fully exploit the abilities of aerial image detectors," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[12] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3578–3587.

[13] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," in *Advances in neural information processing systems*, 2018, pp. 9310–9320.

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

[15] 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)*, 2018, pp. 370–386.

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

[17] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

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

[19] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9657–9666.

[20] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6023–6032.

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[22] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and small object detection in uav vision based on cascade network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[23] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," *arXiv preprint arXiv:1902.07296*, 2019.

[24] C. Chen, Y. Zhang, Q. Lv, S. Wei, X. Wang, X. Sun, and J. Dong, "Rrnet: A hybrid detector for object detection in drone-captured images," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[25] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.