HRDNET: HIGH-RESOLUTION DETECTION NETWORK FOR SMALL OBJECTS

Ziming Liu[*‡], *Guangyu Gao*[*✉], *Lin Sun*[†] *and Zhiyuan Fang*[*]

[*]Beijing Institute of Technology, Beijing, China; [†]Magic Leap, Inc., CA, USA; and
[‡]Inria, Université Côte d'Azur, Sophia Antipolis, France.
liuziming.email@gmail.com; guangyugao@bit.edu.cn; sunlin@cs.stanford.edu; zy_fang@outlook.com

ABSTRACT

Small object detection is a very challenging yet practical vision task. With deep network-based methods, the contextual information of small objects may disappear when the network goes deeper. An intuitive solution to alleviate this issue is to increase the input resolution, however, it will aggravate the large variant of object scale and introduce unbearable computation cost. To leverage the benefits of high-resolution images without bringing up new problems, we propose a High-Resolution Detection Network (HRDNet) which takes multiple resolution inputs with multi-depth backbones. Meanwhile, we propose the Multi-Depth Image Pyramid Network (MD-IPN) and Multi-Scale Feature Pyramid Network (MS-FPN). The MD-IPN maintains multiple position information using multiple depth backbones. Specifically, high-resolution input will be fed into a shallow network to reserve more positional information and reduce computational costs, while low-resolution input will be fed into a deep network to extract more semantics. By extracting various features from high to low resolutions, the MD-IPN can improve the performance of small object detection and maintain the performance of middle and large objects. Additionally, MS-FPN is introduced to align and fuse multi-scale feature groups generated by MD-IPN to reduce the information imbalance. Extensive experiments are conducted on the COCO2017 and the typical small object dataset, VisDrone 2019. Notably, our HRDNet achieves the state-of-the-art on these two datasets with significant improvements on small objects.

***Index Terms*—** Small Object Detection, High-resolution Images, Image Pyramid, Deep Neural Network

1. INTRODUCTION

With the advances of deep learning, object detection achieves the remarkable progress. According to whether the proposals are generated by an independent learning stage or directly and densely sample possible locations, object detection can be classified into two-stage or one-stage models. Compared to two-stage detectors [1, 2], one stage methods [3, 4] are less complex and faster with some precision loss. Recently, lots of anchor-free detectors are proposed, e.g., CornerNet [5], FCOS [6], FSAF [7], which create some low-level abstractions of the images like lines, circles, and then 'iteratively combine' them into some objects. However, these methods are still struggling with small objects.

Those above-mentioned methods can benefit from the deep and powerful network with multi-level features fusion, e.g., using feature pyramid network (FPN), but also it will introduce more computations. While high-resolution images reserve more detail information, they are helpful for the small object detection. However, they also introduce new issues, such as that, it will degrade the performance of large objects and bring in unaffordable computation cost. We focus on the trade-offs between large and small object detection and high performance and low computational complexity.

In this paper, we propose a novel High-resolution Detection Network (HRDNet), which includes a Multi-Depth Image Pyramid Network (MD-IPN) and a Multi-Scale Feature Pyramid Network (MS-FPN), as shown in Figure 1. The core idea of the HRDNet is to use a deep backbone to process low-resolution images while using a shallow backbone to process high-resolution images. The advantage of extracting features from high-resolution images with the shallow and tiny network has been demonstrated in [8]. With HRDNet, we can not only get more details for a small objects from high-resolution images, but also guarantee the efficiency and effectiveness by integrating multi-depth and multi-scale deep networks.

The MD-IPN can be regarded as a variant of the image pyramid network with multiple streams. MD-IPN is dealing with the trade-offs between large and small object detection, as well as high performance and low computational complexity. We extract features from the high-resolution image using a shallow backbone network. Because of its weak representation power, we also need deep backbones to obtain robust semantic features by feeding low-resolution images in. Thus, the inputs of the MD-IPN form an image pyramid with a fixed decreasing ratio of $\alpha \in [0, 1]$. The output of MD-IPN is a series of multi-scale feature groups, while each group contains multi-level feature maps.

The multi-scale feature groups extend the standard feature

✉ Corresponding author. This work was supported mainly by the National Natural Science Foundation of China under Grant 61972036, in part by grants under Grant 91746210.

978-1-6654-3864-3/21/$31.00

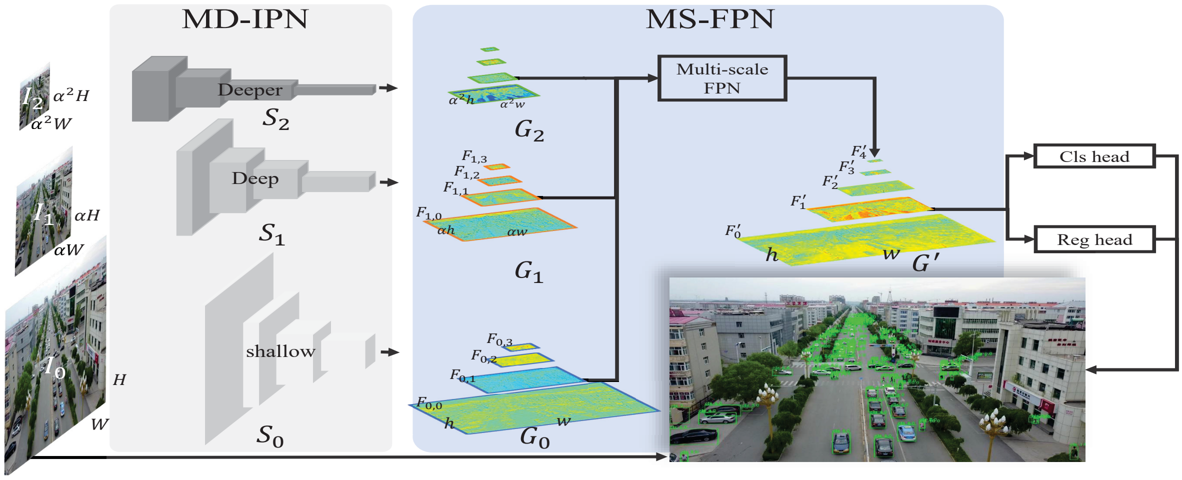


Fig. 1. Overall structure of HRDNet with MD-IPN and MS-FPN. The input is an image pyramid with N images ($N = 3$ here), and the decreasing ratio is α. The outputs of MD-IPN are N groups feature pyramids, and the decreasing ratio of each feature pyramid is 2. MS-FPN fuses these features into one feature pyramid $\{F_0^{'}, F_1^{'}, F_2^{'}, F_3^{'}, F_4^{'}\}$, which is used for object detection.

pyramid by adding multi-scale streams. Therefore, traditional FPN cannot be directly applied here. To fuse these multi-scale feature groups properly, we proposed the Multi-Scale Feature Pyramid Network (MS-FPN). As shown in Figure 2, the information of images not only propagates from high-level features to low-level features inside the multi-level feature pyramid, but also between streams of different scales in MD-IPN.

We summarize our contributions as follows:

- We comprehensively analyzed the factors that small object detection depends on and the trade-off between performance and efficiency and proposed the HRDNet, considering both image pyramid and feature pyramid.

- We designed the multi-depth and multi-stream module, i.e., MD-IPN, to balance the performance between small, middle, and large objects. We also proposed the MS-FPN, to combine different semantic representations from multi-scale feature groups.
- Extensive ablation studies validate the effectiveness and efficiency of the proposed approach when our approach achieves state-of-the-art performance, particularly on small object detection.

2. RELATED WORK

The state-of-the-art object detection methods include one stage models, e.g., RetinaNet [3], Yolo-v3 [9], Center net [10], FSAF [7], Corner net [5], EfficientDet [11] and two-stage models, e.g., Faster R-CNN [2], Cascade R-CNN [1] etc. Nevertheless, our HRDNet is a more fundamental and general framework for most of detection models, such as RetinaNet and Cascade R-CNN.

2.1. Small object detection

The detection performance is largely bounded by small object detection. Therefore, there are many researches specializing in small object detection. For example, [12] proposed oversampling and copy-pasting small objects to solve such a problem. Perceptual GAN [13] generated super-resolved features and stacked them into feature maps of small objects to enhance the representations. DetNet [14] maintained the spatial resolution and has a large receptive field to improve small object detection. SNIP [15] resized images to different resolutions and only train samples which is close to ground truth. SNIPER [16] is proposed to use regions around the bounding box to remove the influence of background. Unlike these methods, we combine both image pyramid and feature pyramid together, with which it not only effectively improves the detection performance of small targets, but also ensure the detection performance of other objects.

2.2. High-resolution detection

Some studies already explored to do object detection on high-resolution images. [8] proposed a fast tiny detection network for high-resolution remote sensing images. [17] proposed an attention pipeline to achieve fast detection on 4K or 8K videos using YOLO v2 [18]. However, these works did not fully explore the effect of high-resolution images for small object detection, which is what we concentrate on.

2.3. Feature-level imbalance

To capture the semantic information of objects from different scales, multi-level features are commonly used for object detection. However, they have serious feature-level imbalance because they convey different semantic information. Feature

Pyramid Network (FPN) [19] introduced a top-down pathway to transmit semantic information, alleviating the feature imbalance problem in some degree. Based on FPN, PANet [20] involved a bottom-up path to enrich the location information of deep layers. The Libra R-CNN [21] revealed and tried to deal with the sample level, feature level, and objective level imbalance issues. Pang et al. [22] proposed a light weighted module to produce featured image pyramid features to augment the output feature pyramid. While these methods only focus on multi-level features, we proposed a new module called Multi-scale FPN to solve the imbalance not only from multi-level features but also from multi-scale feature groups.

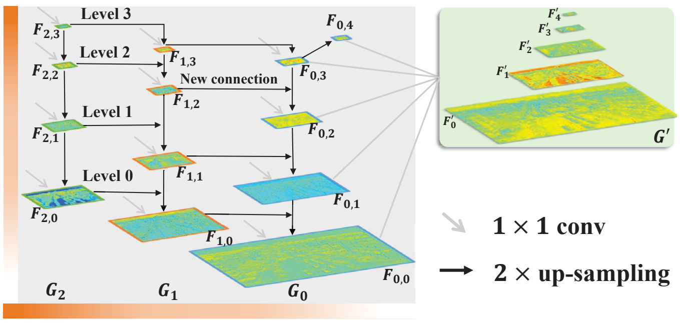


Fig. 2. MS-FPN, a feature pyramid with three streams and four levels. The horizontal orange bar indicates the depth of $\mathcal{S}_i$, the vertical orange bar indicates the depth of a single backbone. Better viewed in color and zoom in.

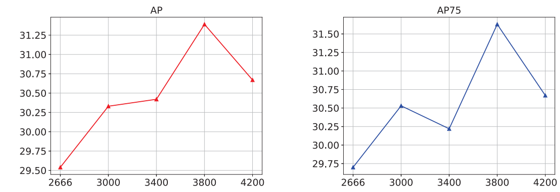
3. HIGH-RESOLUTION DETECTION NETWORK

Intuitively, high-resolution images are good for small object detection. Unfortunately, high-resolution images will introduce unaffordable computation costs to deep networks. Meanwhile, high-resolution images aggravate the variance of object scales, making the performance of large objects worse, as shown in Table 1. To balance computation costs and variance of objects scales while keeping the performance of all the classes, we proposed the High-Resolution Detection Network (HRDNet). The HRDNet is a general concept that is compatible with any alternative detection methods.

Specifically, HRDNet is designed with two modules: Multi-Depth Image Pyramid Network (MD-IPN) and Multi-Scale Feature Pyramid Network (MS-FPN). In MD-IPN, an image pyramid is processed by backbones with different depth, i.e., using deep CNNs for the low-resolution images while using shallow CNNs for the high-resolution images, as shown in Figure 1. After that, to fuse the multi-scale groups of multi-level features from MD-IPN, MS-FPN is proposed as a more reasonable feature pyramid architecture (Figure 2).

3.1. MD-IPN

The MD-IPN is composed of N independent backbones with various depth to process the image pyramid. We term each



(a) (b)

Fig. 3. The change of AP, $AP75$ over different input's resolution. The HRDNet used here is a two-stream version with ResNet18+101 backbone.

backbone as a *stream*. HRDNet can be generalized to more streams, but for better illustration, we mainly discuss the two-stream HRDNet and three-stream HRDNet. Figure 1 presents an example of three-stream HRDNet. Given an image I with resolution R, the high-resolution image (I_0 with R) is processed by a stream of shallow CNN (S_0), and the lower-resolution images (I_1 and I_2 with αR and $\alpha^2 R$, and $\alpha = 0.5$.) is processed by streams of deeper CNN (S_1 and S_2). Generally, we can build an image pyramid network with N independent parallel streams, $\mathcal{S}_i, i = \{0, 1, \ldots, N-1\}$.

We use $\{I_i\}_{i=0}^{N-1}$ to represent the input images with different resolutions given the *original image* I_0 with the highest resolution. The outputs of the multi scale image pyramid are N feature groups $\{\mathcal{G}_i\}_{i=0}^{N-1}$. Each group $\mathcal{G}_i$ contains a set of multi-level features $\{F_{i,j}\}$, where $i \in \{0, 1, \ldots, N-1\}$ is the multi-scale index and $j \in \{0, 1, \ldots, M-1\}$ is the multi-level index. For example, in Figure 1, N and M are 3, 4 respectively, and the relation can be formulated as $\mathcal{G}_i = \mathcal{S}_i(I_i) = \{F_{i,0}, F_{i,1}, F_{i,2}, F_{i,3}\}$, where $i \in \{0, 1, \ldots, N-1\}$.

3.2. MS-FPN

Feature pyramid network (FPN) is one of the key components for most object detection algorithms. It combines low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections. In our HRDNet, the MD-IPN generates multi-scale (different resolution) and multi-level (different hierarchy of features) features. To deal with the multi-scale hierarchy features, we also proposed the Multi-Scale FPN (MS-FPN). Different from FPN, semantic information propagates not only from high-level features to low-level features but also from deep stream (low-resolution) to shallow stream (high-resolution). Therefore, there are two directions for the computation of the multi-scale FPN. The basic operation in multi-scale FPN is same as traditional FPN, i.e., 1×1Convolution, $2\times$ up-sampling and sum-up.

In this way, the highest resolution feature, i.e., $F_{0,0}$, not only maintain the high-resolution for small object detection but also combine semantically strong features from multi-scale streams. Our novel MS-FPN can be formulated as $F_{i,j} = Conv(F_{i,j}) + Up(F_{i,j+1})$ if $i = N$-1, $F_{i,j} = Conv(F_{i,j}) + Up(F_{i,j+1}) + Up(F_{i+1,j})$ if $i \neq N$-1.

Table 1. **Performance Comparison of Cascade R-CNN and HRDNet with different resolution's input.** The HRDNet here is a two streams version, and † means that it is trained on patch images as mentioned in Section 4.1.

model	resolution	pedestrian	people	bicycle	car	van	truck	tricycle	awning-tri	bus	motor	**mAP**
Cascade R-CNN	1333×800	37.9	27.7	13.3	74.3	44.6	34.7	24.6	13.2	52.4	38.3	**36.1**
Cascade R-CNN	2666×1600	51.5	38.0	20.2	80.0	48.0	32.4$\Downarrow$	28.2	12.1$\Downarrow$	44.8$\Downarrow$	47.5	**40.3**
HRDNet	2000×1200	49.6	37.2	17.4	79.8	47.9	36.9	30.4	15.3	56.0	48.7	**41.9**
HRDNet	2666×1600	55.8	42.4	23.1	82.4	51.2	42.1$\Uparrow$	34.3	16.3$\Uparrow$	59.7$\Uparrow$	53.8	**46.1**
HRDNet †	2666×1600	56.7	45.1	27.7	82.6	51.3	43.0	37.6	18.8	58.9	56.4	**47.8$\Uparrow$**

The $F_{i,j}$ is the feature in level j and stream i in Figure 2. The $Up(.)$ operation is $2\times$ up-sampling. The $Conv(.)$ is 1×1 convolution. Finally, MS-FPN outputs the final feature group $G' = \{F'_0, F'_1, ...F'_i, ...\}$. F'_i is calculated by $F'_i = Conv(F_{0,i})$, where $F_{0,i}$ is the features in Group G_0, i.e., the outputs of the highest resolution stream.

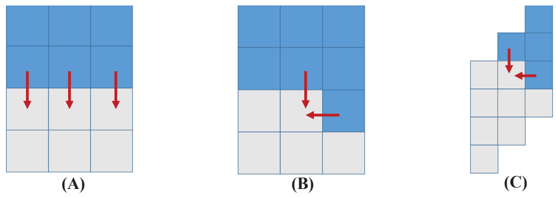


Fig. 4. Comparison of the simple FPN (A), multi-scale FPN aligned with depth (B) and resolution (C). Each column is one stream in MD-IPN, and each row refers to the depth of backbone. The blue and gray blocks are those features been fused and features to be fused respectively. The red arrows is a basic fusing operation described in subsection 2.3.

4. EXPERIMENTS

4.1. Experiment details

Datasets. We conduct experiments on two typical small object detection data sets: VisDrone2019 [23] and COCO2017 [24]. The VisDrone2019 dataset consists of 288 video clips formed by 261,908 frames and 10,209 static images, covering a wide range location, environment, objects, and density. The resolution of VisDrone2019 is ranging from 960 to 1360. COCO is the most common benchmark for object detection, and we trained our model on the COCO training set and tested it on the COCO validation set. In COCO, most images' resolution is 500-800 px, which will be resized to 1333×800 or 1000×600 in the training stage, but 960-1360 px in VisDrone2019 [25] dataset.

Training. We use SGD optimizer with a mini-batch 2 for each GPU. The learning rate starts from 0.02 and decreases by 10 at epoch 7 and 11 with totally 15 epochs. The weight decay is 1×10^{-4}. The linear warm-up strategy is used with warm-up iterations of 500, and the warm-up ratio of $1.0/3$. The linear interpolation is applied for the image pyramid. The resolution decreasing ratio α is 0.5. To fit the high-resolution images from VisDrone2019 into the GPU memory, we equally cropped each original image in VisDrone2019 training set into four patches which are not overlapped. In this way, we obtained a new training set with such cropped images.

Inference. Same resolution as training is used for inference. The NMS IOU threshold is 0.5, and the threshold of confidence score is 0.05. Without especially emphasizing, three scales are applied for the multi-scale test.

4.2. Ablation Studies

The effect of image resolution. Extensive ablation studies on the VisDrone2019 dataset are conducted to illustrate the effect of input image resolution for detection performance. Table 1 shows that the performance has a significant improvement with the increase of image resolution. Higher resolution leads to better performance under the same experimental settings. The detection of small objects show more improvements. Importantly, HRDNet performs much better than the SOTA Cascade R-CNN with the same resolution as the input.

Interestingly, when the resolution of input increases, single backbone model, i.e., Cascade R-CNN, suffers dramatically decrease (1.1-7.6%) for categories with relatively large size, i.e., *truck*, *awning-tricycle* and *bus*. On the contrary, significant performance increase (1-5.2%) can be observed from HRDNet. Simply increasing the image resolution without considering the severe variant of object scale is not the ideal solution for detection, let alone small object detection.

Explore the optimal image resolution. Is it true higher resolution leads to better performance? Does it have the optimal resolution for detection? In this part, we will present the effect of image resolution for object detection. Figure 3 shows the change of the Average Precise ($AP[0.05:0.95]$, $AP75$) with different resolutions. The resolution starts from 2666 (long edge) with 400 as the stride. HRDNet achieves the best performance when the resolution is 3800×2800 px.

Table 2. The comparison of three different MS-FPNs.

style	AP	AP50	AP75
ResNet10+18			
simple FPN	28.8	49.5	28.8
aligned by resolution	28.7	49.6	28.7
aligned by depth	**28.9**	**49.9**	28.7
ResNet18+101			
aligned by resolution	31.8	54.0	32.3
aligned by depth	**32.0**	**54.3**	**32.5**

Table 3. The speed (items/s) and the number of parameters (M) are obtained on a same machine with one *Nvidia GTX 2080Ti GPU* and *Intel(R) Xeon(R) Silver 4210 CPU*. Here is a two stream HRDNet using MS-FPN aligned with depth.

model	backbone	resolution	params	speed	AP50
Cascade	ResNet18	1333	56.11M	9.9	36.1
Cascade	ResNet18	2666	56.11M	5.4	40.3
Cascade	ResNet18	3800	56.11M	2.9	42.6
HRDNet	ResNet10+18	3800	62.44M	3.7	49.2
HRDNet	ResNet18+101	3800	100.78M	2.8	53.3
HRDNet	ResNeXt50+101	3800	152.22M	1.5	55.2

Table 4. The comparison of HRDNet and Model Ensemble. The models here follow the design of Cascade R-CNN.

model	backbone	resolution	AP	AP50	AP75
Single Backbone	ResNeXt50	3800	32.7	54.6	33.6
Single Backbone	ResNeXt101	1900	30.4	51.0	31.1
Model Ensemble	ResNeXt50+101	3800+1900	32.9	55.1	33.5
HRDNet	ResNeXt50+101	3800+1900	**33.5**	**56.3**	**34.0**

Table 5. The comparison with SOTA models on visdrone2019 DET validation set. † results are obtained with the same environment. ++ denotes multi-scale test.

model	backbone	resolution	AP	AP50	AP75
†Cascade R-CNN [1]	ResNet50	2666	24.10	42.90	23.60
†Faster R-CNN [2]	ResNet50	2666	23.50	43.70	22.20
†RetinaNet[3]	ResNet50	2666	15.10	27.70	14.30
†FCOS [6]	ResNet50	2666	16.60	28.80	16.70
HFEA [26]	ResNeXt101	-	27.10	-	-
HFEA [26]	ResNeXt152	-	30.30	-	-
DSOD [27]	ResNet50	-	28.80	47.10	29.30
†HRDNet	ResNet10+18	3800	28.68	49.15	28.90
†HRDNet	ResNet18+101	2666	28.33	49.25	28.16
†HRDNet	ResNet18+101	3800	31.39	53.29	31.63
†HRDNet++	ResNet50+101+152	3800	34.35	56.65	35.51
†HRDNet++	ResNeXt50+101	3800	**35.51**	**62.00**	**35.13**

Table 6. The state of the art performance on COCO *test-dev*, the input resolution of HRDNet ResNet101 stream is same as other models above, i.e. 1333×800, while the input of ResNet 152 stream is a $2\times$ smaller image. '++' denotes that the inference is performed with multi-scales.

model	backbone	AP	AP_S	AP_M	AP_L
Faster R-CNN w FPN [19]	ResNet-101	36.2	18.2	39.0	48.2
DeNet-101(wide) [28]	ResNet-101	33.8	12.3	36.1	50.8
CoupleNet [29]	ResNet-101	34.4	13.4	38.1	50.8
Mask-RCNN [30]	ResNeXt-101	39.8	22.1	43.2	51.2
Cascade RCNN [1]	ResNet-101	42.8	23.7	45.5	55.2
SNIP++ [15]	ResNet-101	43.4	27.2	46.5	54.9
SNIPER(2scale) [16]	ResNet-101	43.3	27.1	44.7	56.1
Grid-RCNN [31]	ResNeXt-101	43.2	25.1	46.5	55.2
SSD512 [4]	VGG-16	28.8	10.9	31.8	43.5
RetinaNet80++ [3]	ResNet-101	39.1	21.8	42.7	50.2
RefineDet512 [32]	ResNet-101	36.4	16.6	39.9	51.4
M2Det800	VGG-16	41.0	22.1	46.5	53.8
CornetNet511 [5]	Hourglass-104	40.5	19.4	42.7	53.9
FCOS [6]	ResNeXt-101	42.1	25.6	44.9	52.0
FSAF [7]	ResNeXt-101	42.9	26.6	46.2	52.7
CenterNet511 [10]	Hourglass-104	44.9	25.6	47.4	57.4
HRDNet++	ResNet101+152	**47.4**	**32.1**	**50.5**	**55.8**

How to design the multi-scale FPN. MS-FPN is designed to fuse multi-scale feature groups. Here, we consider three different styles, including *simple FPN, multi-scale FPN aligned by depth, multi-scale FPN aligned by resolution*, as shown in Figure 4, to demonstrate MS-FPN's advantage. A simple FPN is to apply standard FPN to each multi-scale group of HRDNet and fuse the results of each FPN. New connections between multi-streams are introduced for multi-scale FPN, as shown in Figure 2. We conducted two groups experiments with ResNet10+18 and ResNet18+101 backbone. The first experiment in Table 2 shows that the multi-scale FPN is better than the simple FPN. Both experiments demonstrate that MS-FPN aligned with depth performs better than those aligned with resolution. Therefore, we adopt MS-FPN aligned with depth in our architecture.

Efficient and Effective HRDNet. We illustrate the number of parameters and running speed of proposed HRDNet, comparing with the state-of-the-art single backbone baseline. The results are shown in Table 3 prove that HRDNet can achieve much better performance with a similar number of parameters and even faster running speed.

Comparison with model ensemble. To further demonstrate that the improvement of HRDNet *is not due to more parameters*, we compared a two-stream HRDNet with the ensemble of two backbone models under the same experimental setting (Table 4). The ensembled models fuse the predicted bounding boxes and scores before NMS (Non-Maximum Suppression) and then perform NMS. We found that the single backbone models with high-resolution input always perform better than those with low-resolution even it is processed by a stronger backbone. HRDNet performs better than the ensemble model, thanks to the novel multi-scale and multi-level fusion method. They further prove that our designed MS-FPN is essential for HRDNet.

4.3. Comparison with the state-of-the-art methods

VisDrone2019: We compare HRDNet with the SOTA methods to demonstrate the advantage of our model and technical contributions. Table 5 shows that HRDNet achieves the best performance on VisDrone2019 DET validation set. Notably, our model obtains more than 3.0% AP improvement with ResNeXt50+101 compared to HFEA (ResNet152).
COCO2017: Besides the experiments on VisDrone2019, we also conduct experiments on the COCO2017 test set to prove our method works well on a larger scale, complicated and standard detection dataset. Table 6 shows that HRDNet achieves state-of-the-art results, and $> 4.9\% AP_{small}$ improvement compared with most recent algorithms.

5. CONCLUSION

In this paper, we propose the HRDNet with well-designed MD-IPN and MS-FPN to solve the issues which are not considered in others for small objects. HRDNet achieves the state-of-the-art on small object detection dataset, VisDrone2019, at the same time, we outperform with a large margin on COCO.

6. REFERENCES

[1] Zhaowei Cai and Nuno Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *CVPR*, 2018, pp. 6154–6162.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Trans. on PAMI*, vol. 39, no. 6, pp. 1137–1149, 2016.

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: single shot multibox detector," in *ECCV*, 2016, pp. 21–37.

[5] Hei Law and Jia Deng, "Cornernet: Detecting objects as paired keypoints," in *ECCV*, 2018, pp. 765–781.

[6] Zhi Tian, Chunhua Shen, and et al., "FCOS: Fully convolutional one-stage object detection," in *ICCV*, 2019, pp. 9626–9635.

[7] Chenchen Zhu, Yihui He, and Marios Savvides, "Feature selective anchor-free module for single-shot object detection," in *CVPR*, 2019, pp. 840–849.

[8] ," .

[9] Joseph Redmon Ali Farhadi and Joseph Redmon, "Yolov3: An incremental improvement," *arXiv*, 2018.

[10] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian, "Centernet: Keypoint triplets for object detection," in *ICCV*, 2019, pp. 6569–6578.

[11] Mingxing Tan, Ruoming Pang, and Quoc V Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10781–10790.

[12] Mate Kisantal, Zbigniew Wojna, Jakub Murawski, Jacek Naruniec, and Kyunghyun Cho, "Augmentation for small object detection," *arXiv preprint arXiv:1902.07296*, 2019.

[13] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan, "Perceptual generative adversarial networks for small object detection," in *CVPR*, 2017.

[14] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun, "Detnet: A backbone network for object detection," *arXiv*, 2018.

[15] Bharat Singh and Larry S Davis, "An analysis of scale invariance in object detection snip," in *CVPR*, 2018, pp. 3578–3587.

[16] Bharat Singh, Mahyar Najibi, and et al., "SNIPER: efficient multi-scale training," in *NeurIPS*, 2018.

[17] Vít Ržička and Franz Franchetti, "Fast and accurate object detection in high resolution 4k and 8k video using gpus," in *HPEC*, 2018, pp. 1–7.

[18] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *CVPR*, 2017, pp. 7263–7271.

[19] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.

[20] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *CVPR*, 2018, pp. 8759–8768.

[21] Jiangmiao Pang and Kai Chen, "Libra R-CNN: towards balanced learning for object detection," in *CVPR*, 2019, pp. 821–830.

[22] Yanwei Pang, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao, "Efficient featurized image pyramid network for single shot detector," in *CVPR*, 2019, pp. 7336–7344.

[23] Pengfei Zhu, Dawei Du, and et al., "Visdrone-vid2019: The vision meets drone object detection in video challenge results," in *ICCV Workshop*, 2019.

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, and et al., "Microsoft COCO: common objects in context," in *ECCV*, 2014, pp. 740–755.

[25] Dawei Du, Pengfei Zhu, and et al., "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *ICCV Workshop*, 2019.

[26] Junyi Zhang, Junying Huang, Xuankun Chen, and Dongyu Zhang, "How to fully exploit the abilities of aerial image detectors," in *ICCV Workshop*, 2019.

[27] Xindi Zhang, Ebroul Izquierdo, and Krishna Chandramouli, "Dense and small object detection in uav vision based on cascade network," in *ICCV Workshop*, 2019, pp. 118–126.

[28] Amir Ghodrati, Ali Diba, Marco Pedersoli, Tinne Tuytelaars, and Luc Van Gool, "Deepproposal: Hunting objects by cascading deep convolutional layers," in *ICCV*, 2015, pp. 2578–2586.

[29] Yousong Zhu, Chaoyang Zhao, Jinqiao Wang, Xu Zhao, Yi Wu, and Hanqing Lu, "Couplenet: Coupling global structure with local parts for object detection," in *ICCV*, 2017, pp. 4126–4134.

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.

[31] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan, "Grid r-cnn," in *CVPR*, 2019, pp. 7363–7372.

[32] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li, "Single-shot refinement neural network for object detection," in *CVPR*, 2018, pp. 4203–4212.