# $S^3OD$: Single Stage Small Object Detector from Scratch for Remote Sensing Images

Feng Yang[1,2,4(✉)], Wentong Li[1,2], Wanyi Li[3], and Peng Wang[3]

[1] Northwestern Polytechnical University,
Xi'an 710129, People's Republic of China
yangfeng@nwpu.edu.cn
[2] Key Laboratory of Information Fusion Technology, Ministry of Education,
Xi'an 710129, People's Republic of China
[3] Institute of Automation, Chinese Academy of Sciences, Beijing 100190,
People's Republic of China
[4] CETC Key Laboratory of Data Link Technology,
Xi'an 710129, People's Republic of China

**Abstract.** Small object detection is an important but challenge computer vision task in both natural scene and remote sensing scene. Due to the large difference of density, low contrast, sparse texture and arbitrary orientations, many advanced algorithms for small object detection in natural scene usually experience a sharp performance drop when directly applied to remote sensing images. In addition, most of state-of-the-art object detectors are fine-tuned from the off-the-shelf networks pretrained on large-scale classification dataset like ImageNet, which can incur learning bias and inconvenience of modification for remote sensing object detection tasks. In order to tackle these problems, a robust Single Stage Small Object Detector ($S^3OD$) is trained from scratch, which can efficiently detect small-dense and small-dispersed objects in remote sensing images. The proposed $S^3OD$ adopts the small down-sampling factor to keep accurate location information and maintains high spatial resolution by introducing a new dilated residual block in deeper layers for small objects. Especially, the two-branch dilated feature attention module is proposed to enlarge the valid receptive field and make effective attention feature map for small-dense and small-dispersed object detection. $S^3OD$ can be trained from scratch stably while keeping the comparable performance by employing BatchNorm on both the backbone and detection head subnetworks. Experiments conducted on our built Remoting Sensing Small Object (RSSO) dataset shows that, our $S^3OD$ achieves the state-of-the-art accuracy for small objects detection and even performs better than several one-stage pretrained method.

**Keywords:** Object detection · Remote sensing images · Small objects · Convolutional neural networks


Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11903, pp. 347–358, 2019.
https://doi.org/10.1007/978-3-030-34113-8_29

## 1 Introduction

Object detection of remote sensing images plays an important role in many real-world applications such as traffic control, environmental monitoring, and urban planning. Remarkable progresses have been made in object detection of remoting sensing images recently due to the convolutional neural networks (CNNs) [1]. However, the small object detection is still one of remaining challenge tasks in remote sensing images [2]. Lots of CNN-based object detectors have been proposed and achieve great success over natural scene. Those methods can be divided into two categories including one-stage detector, like YOLO [3], SSD [4], and two-stage detectors, like Faster R-CNN [5] and R-FCN [6]. It is found experimentally that these frameworks have poor performance for small objects, because they are based on high-level CNN features and fail to capture precise descriptions of small objects. For more advanced methods, FPN [7] introduces feature pyramids to combine multi-layer feature map by utilizing U-shape structure. RetinaNet [8] proposes a new focal loss to address class imbalance issue to make the object detection more accurate. YOLOv3 [9] introduces a powerful feature extraction backbone and adopts a similar concept to feature pyramid networks in detection layers. TridentNet [10] constructs a parallel multi-branch architecture with different receptive fields to detection multi-scale object. Those frameworks have achieved promising results for small objects detection to a certain extent. However, these methods often experience a sharp performance drop while directly applied to remote sensing images to detect small objects. The main reasons are as followed and illustrated in Fig. 1.

(1) Small objects in remote sensing images usually appear in dense cluster with overwhelmed feature information or in dispersed distribution with sparse feature information.
(2) Remote sensing objects viewed from over-head appear in arbitrary orientations. Such as the ship can have any degrees between 0 and 360 degrees, whereas the objects in ImageNet are often vertical.
(3) Remote sensing images are complex, not only because a large amount of various noises from remote sensors (like satellite sensors), but also because remote sensing objects usually lack visual clues such as texture details, image contrast.

For remote sensing images, many deep-learning-based detection methods also have been developed. R2-CNN [11] proposes a unified and self-reinforced network including Tiny-Net backbone, global attention block and final classifier and detector towards practical real-time remote sensing systems. YOLT [12] inspired by YOLOv2 [3] implements a unique network architecture with a denser final prediction grid to help differentiate between classes by yielding grained features in remote sensing images. Those methods mainly focus on how to implement a multi-class framework elegantly while the detection performance of small objects is not well. Small object detection seems much more difficult.

Besides, most of current impressive detectors are generally fine-tuned from the off-the-shelf networks with high accuracy classification, e.g. VGGNet [13], ResNet [14] pretrained on ImageNet dataset. Object detectors fine-tuned from pretrained networks often achieve better performance than those trained from scratch. But fine-tuning from

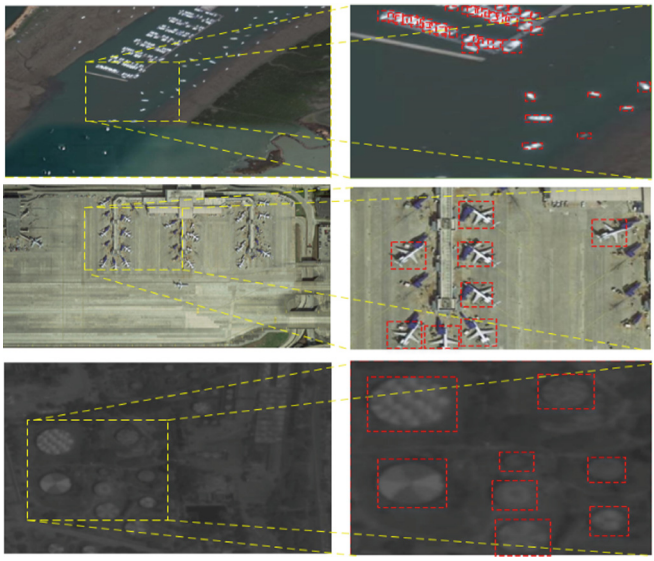

**Fig. 1.** Examples of small object detection including boat, airplane and oilcan. Illustration of density, sparse texture, arbitrary orientations, and low contrast in remote sensing images.

pretrained networks for object detection tasks has two main problems: (1) The classification and detection have different loss function, leading to the learning bias; (2) The architecture of backbone is limited by the classification network, resulting in the inconvenience of modification. DSOD [15] is the first to train the one-stage object detector from scratch and focuses on the deep supervision of DenseNet [16]. That introduces many principles to get the good performance. DetNet [17] analyzes the drawbacks of ImageNet pre-trained model for fine-tuning object detectors and presents a train-from-scratch backbone for object detection. ScratchDet [18] explores that BatchNorm is one of the key points for object detectors from scratch and presents a single-shot object detector which integrates BatchNorm to help the detector converge well from scratch. For object detection in remote sensing images, most state-of-art methods are fine-tuned from the pretrained on large-scale dataset ImageNet, which is unreasonable.

In this paper, we propose a Single Stage Small Object Detector (S$^3$OD) for small object detection from scratch in remote sensing images. Firstly, a novel backbone aimed at small remote sensing objects is designed. A large down-sampling factor used by most classic methods with the down-sampling operations (e.g. max-pooling and convolution with stride 2) is not a reasonable option in remote sensing images with high resolution. We adopt a small down-sampling factor to keep more precise feature for small objects. To build a deep neural network which can maintain high resolution feature maps in deeper layer, we introduce a new dilated residual block structure. Secondly, we find out that small remote sensing objects can be divided into two categories, including small-dense objects and small-dispersed objects. To detect small objects effectively, we adopt a two-branch dilated feature attention module, one branch

is designed for small-dense objects with the small dilatation rate in the relatively shallow layer, another one branch is designed for small-dispersed objects with the large dilatation rate in the relatively deep layer. In addition, as pointed out in [20], BatchNorm is one of the key points in current trained-from-scratch detector. We integrate BatchNorm into both the backbone and detection head subnet which helps the detector converge well and achieves the comparable performance without the pretrained baseline.

The main contributions of this paper are summarized as follows. (1) A novel Single Stage Small Object Detector dubbed $S^3$OD is proposed to detect the small objects in remoting sensing images, in which a small down-sampling factor is adopted to keep accurate location information and a new dilated residual block is introduced in deeper layers to maintain high spatial resolution feature maps. (2) We propose to categorize small remote sensing objects into small-dense objects and small-dispersed objects. And a two-branch dilated feature attention module is designed, in which the first branch with small dilatation rate in the relatively shallow layer is for small-dense objects, while the other branch with the large dilatation rate in the relatively deep layer is for small-dispersed objects. (3) To help the $S^3$OD converge well in the train-from-scratch process, the BathchNorm strategy is integrated in each convolutional layer. (4) A Remoting Sensing Small Object (RSSO) dataset is built, and extensive experiments conducted on it demonstrated that our proposed $S^3$OD achieves the state-of-the-art accuracy for small objects detection and even performs better than several one-stage pretrained methods.

## 2 Proposed Method

The overview framework of our proposed Single Stage Small Object Detector ($S^3$OD) is illustrated in Fig. 2. As the figure shown, our $S^3$OD is built on the structure of classic one-stage detection network-YOLOv3. The fine-gained feature map for small-scale objects is extracted by the designed backbone with the down-sampling factor of 16, instead of 32 in standard darknet53 network. Meanwhile the backbone employs several dilated residual blocks to enlarge the receptive filed with the high spatial resolution. The two-branch feature attention module is introduced for small-dense and small-dispersed objects detection in remote sensing images. Finally, BatchNorm is adopted to the whole designed network to train from scratch in a good convergence performance.

### 2.1 Small Down-Sampling Factor in $S^3$OD Backbone

Most of remote sensing object detection methods usually rely on backbone networks like VGGNet [13], ResNet [14], which are used to classification task. The task of classification is different from the object detection which not only needs to recognize the classes of objects but also needs to get the accurate bounding-boxes. Notice that the down-sampling operations (e.g. max-pooling and convolution with stride 2) are one keys of things for translation invariance. In contrast, the local texture information is more critical for object detection especially for complex remote sensing images. In this case, we analyze the performance of VGGNet, ResNet and Darknet53 with various

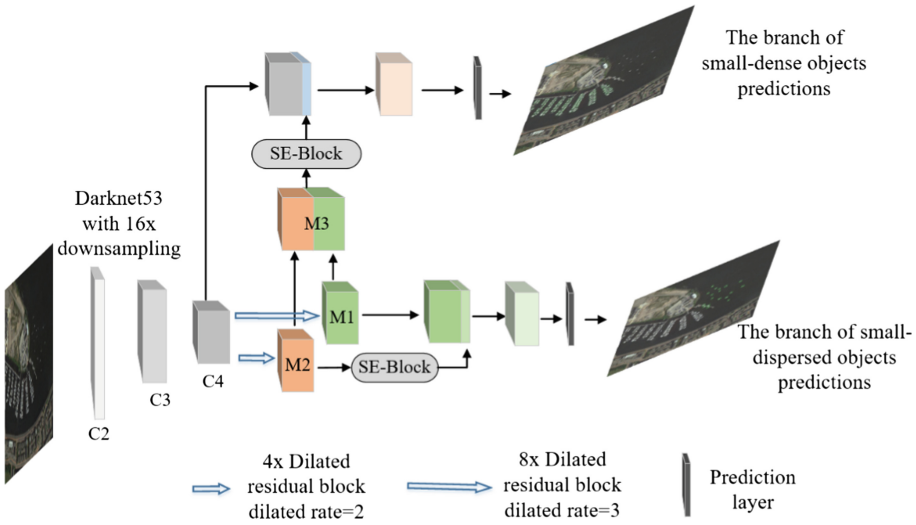


**Fig. 2.** The overview framework of our S$^3$OD method for small object detection in remote sensing images. The backbone is darknet53 with 16x down-sampling factor and the two-branch dilated feature attention module is designed for small-dense objects and small-dispersed objects.

configurations, and discover that the down-sampling factor has a great impact on detection performance. Based on this point, we redesign the architecture of detector by adopting the down-sampling factor of 16. As Fig. 2 shown, the feature map C4, M1 and M2 have the same size of width and height in different layer with 16x down-sampling factor, which keeps the abundant information for detection feature maps and substantially improves the detection accuracy for small objects in remote sensing images.

### 2.2 The Two-Branch Dilated Feature Attention Module

Reducing the down-sampling factor equals to reducing the valid receptive field, which will be harmful for vision tasks. To efficiently enlarge the receptive field, a new dilated residual block structure, which consists of a $1 \times 1$ convolution and a $3 \times 3$ dilated convolution, is adopted to S$^3$OD. Notice that, a dilated $3 \times 3$ convolution with $d_s$ dilation could have the same receptive field as the convolution with kernel size of $3 + 2\ (d_s - 1)$. In additions, Shallow layers usually only have low semantic information which may be not enough to recognize the category of the object instances. Therefore, the 8x dilated residual blocks with dilated rate of 3 and the 4x dilated residual blocks with dilated rate of 2 are constructed to get different receptive field and different depth-level feature map. It is illustrated in Fig. 3.

As Fig. 3 shown, the feature map passed through the 8x dilated residual block from C4 is denoted as M1, the feature map passed through the 4x dilated residual blocks from C4 is denoted as M2. M1 can bring large valid receptive field and high-level representations with a deeper layer, which is mainly designed for the dispersed objects

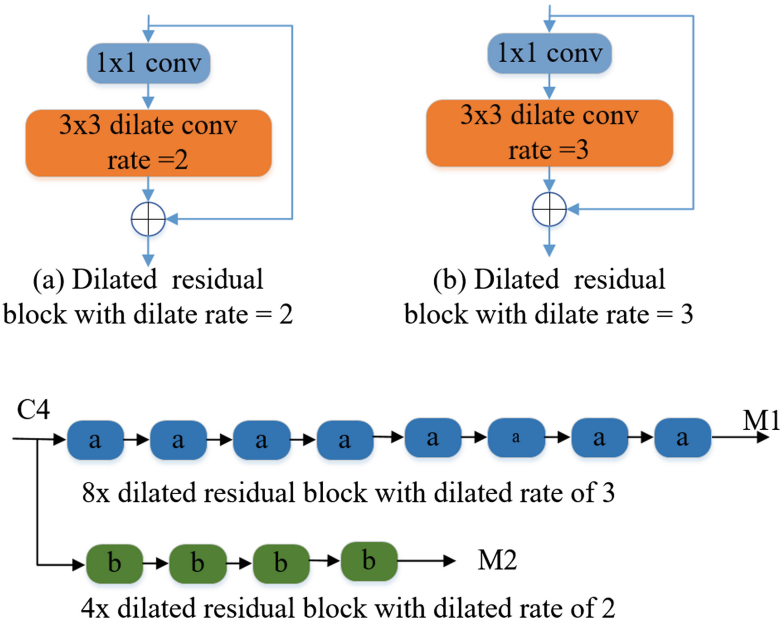


**Fig. 3.** The illustration of different dilated residual block in S$^3$OD in a and b. C4 passes through 8x dilated residual block to get M1, and passes through 4x dilated residual block to get M2 in the two-branch dilated feature attention module.

in remote sensing images. Compared with M1, M2 brings a smaller valid receptive field from a shallower layer, which can make up the loss of information of M1 owing to convolution operation with a lager dilate rate. For small-dense objects detection, we concatenate M1 with M2 to get a high-representation feature map M3 with different depth and different receptive field. C4 feature map keeps the abundant texture information. In order to get the fine-gained feature map, we concatenate C4 with the M3 to detect small-dense objects in remote sensing images. In this two-branch module, we use SE block [20]. SE block can enhance informative features according to attention mechanism and suppress features that are of little use of the current task. Especially for remote sensing images, SE block can weaken the noise and relatively enhance the object attention information. The overview pipeline is shown in Fig. 2.

### 2.3 Training S$^3$OD from Scratch with BatchNorm

BatchNorm can reparameterize the optimization problem to make its landscape significantly smoother instead of reducing the internal covariate shift. We add BatchNorm in each convolution layer in both the backbone and detection head subnetworks, which introduces a more predictable and stable behavior of the great gradients to allow for larger searching space and faster convergence (see Fig. 4). Our proposed train-from-scratch S$^3$OD performs better than several one-stage pretrained models.

## 3 Experiments

In this section, we conduct experiments on our built small dataset of remote sensing images to demonstrate the effectiveness of our proposed S$^3$OD method. The dataset description, implementation details, evaluation metrics, and experimental results will be introduced in detail.

### 3.1 Dataset Description

The proposed method is evaluated over the small object images which are collected from two publicly available datasets including NWPU-VHR10 [21] and AIIA2018_2nd [22]. NWPU-VHR10 has 800 high-resolution remote sensing images in total with 10 classes of objects including plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge and vehicle. The AIIA2018_2nd dataset of remote sensing images is provided from the second stage of AIIA[1] Cup Competition of Typical Object Recognition for Satellite Imagery, which covers six classes: airport, airplane, harbor, boat, oilcan, bridge. The dataset includes 2421 images whose size varies from 512 × 512 pixels to 5120 × 3584.

We select the small objects from the two publicly available datasets above to build a new Remoting Sensing Small Object (RSSO) dataset. RSSO dataset has 1369 images for train and 307 images for test and includes three classes like airplane, ship/boat and storage tank/oilcan which are the common categories in both NWPU-VHR10 and AIIA2018_2nd. Evaluating the images in RSSO, it can be seen that mainly objects are small-dense and small-dispersed and the size of remote sensing object is so small to 5x4, which is a great challenge for small object detection. Some examples of RSSO are given in Fig. 1.

### 3.2 Implementation Details and Evaluation Metrics

The proposed S$^3$OD is trained with Stochastic Gradient Descent (SGD), where momentum is 0.9, the learning rate is 0.01 on a single NVIDIA GeForce GTX 1080Ti GPU with 11 GB memory, along with the deep learning framework PyTorch. Batch size is set to 4. Total training iterations for RSSO dataset are 400 epochs, i.e. 136800 steps. Mean Average Precision is used as the evaluation metric followed by the standard PASCAL VOC criteria, i.e. IoU > 0.5 between ground truths and predicted boxes [23].

### 3.3 Experimental Results

In our experiments, we trained the state-of-the-art algorithms, like YOLOV3, TridentNet, YOLT models with hyper parameter architecture for the purpose of comparison. Our proposed S$^3$OD uses BatchNorm on every convolution layer and train it from scratch. In addition, YOLT and TridentNet are both trained by the way of

[1] AIIA is China Artificial Intelligence Industry Development Alliance.

fine-tuning from the pretrained backbone models. YOLT is trained by the Darknet19 baseline and TridentNet is trained by the Darknet53 baseline. Experimental results on the test set of RSSO dataset are shown in the Table 1. As Table 1 shown, our proposed $S^3$OD outperforms YOLT, YOLOv3 fine-tuned with the pretrained model and TridentNet by 6.8%, 4.5% and 4.1% respectively which demonstrates its effectiveness for small object detection from remote sensing images. Especially for small-dense objects mainly including boat and oilcan, our $S^3$OD has a large improvement because of our introduced two-branch attention module.

**Table 1.** Detection results of YOLOv3, YOLT, TridentNet and our proposed $S^3$OD on mAP over the RSSO test set.

| Method | mAP | Airplane | Boat | Oilcan |
|---|---|---|---|---|
| YOLOv3 without pretrained model | 0.531 | 0.7261 | 0.3055 | 0.5614 |
| YOLT | 0.595 | 0.8903 | 0.3080 | 0.5881 |
| YOLOv3 fine-tuned with pretrained model | 0.618 | 0.8700 | 0.4007 | 0.5840 |
| TridentNet | 0.622 | 0.8644 | 0.4102 | 0.5901 |
| $S^3$OD (ours) | 0.663 | 0.8965 | 0.4667 | 0.6246 |

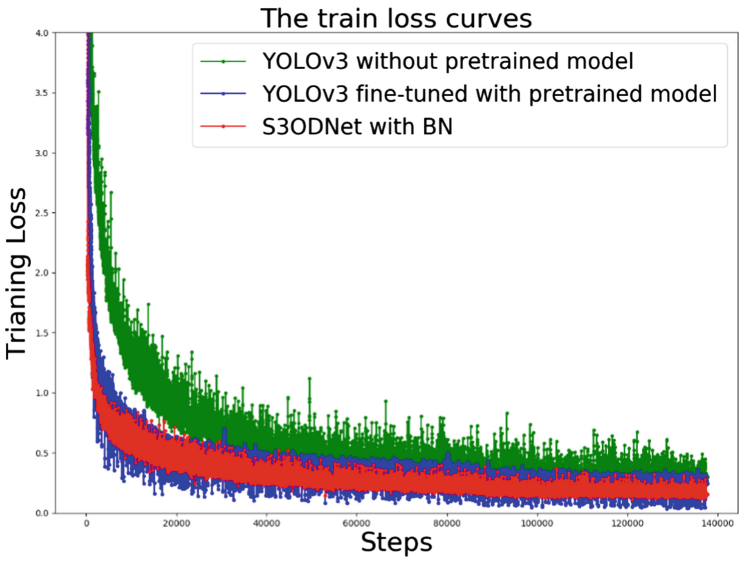


**Fig. 4.** The training loss value is illustrated in this figure. The total loss including x + y coordinates loss, w + h coordinates loss, confidence loss and class score loss are shown. Green and blue curves present YOLOv3 without pretrained model and YOLOv3 fine-tuned with pretrained model respectively, red curve is the train-from-scratch $S^3$OD with BatchNorm. (Color figure online)

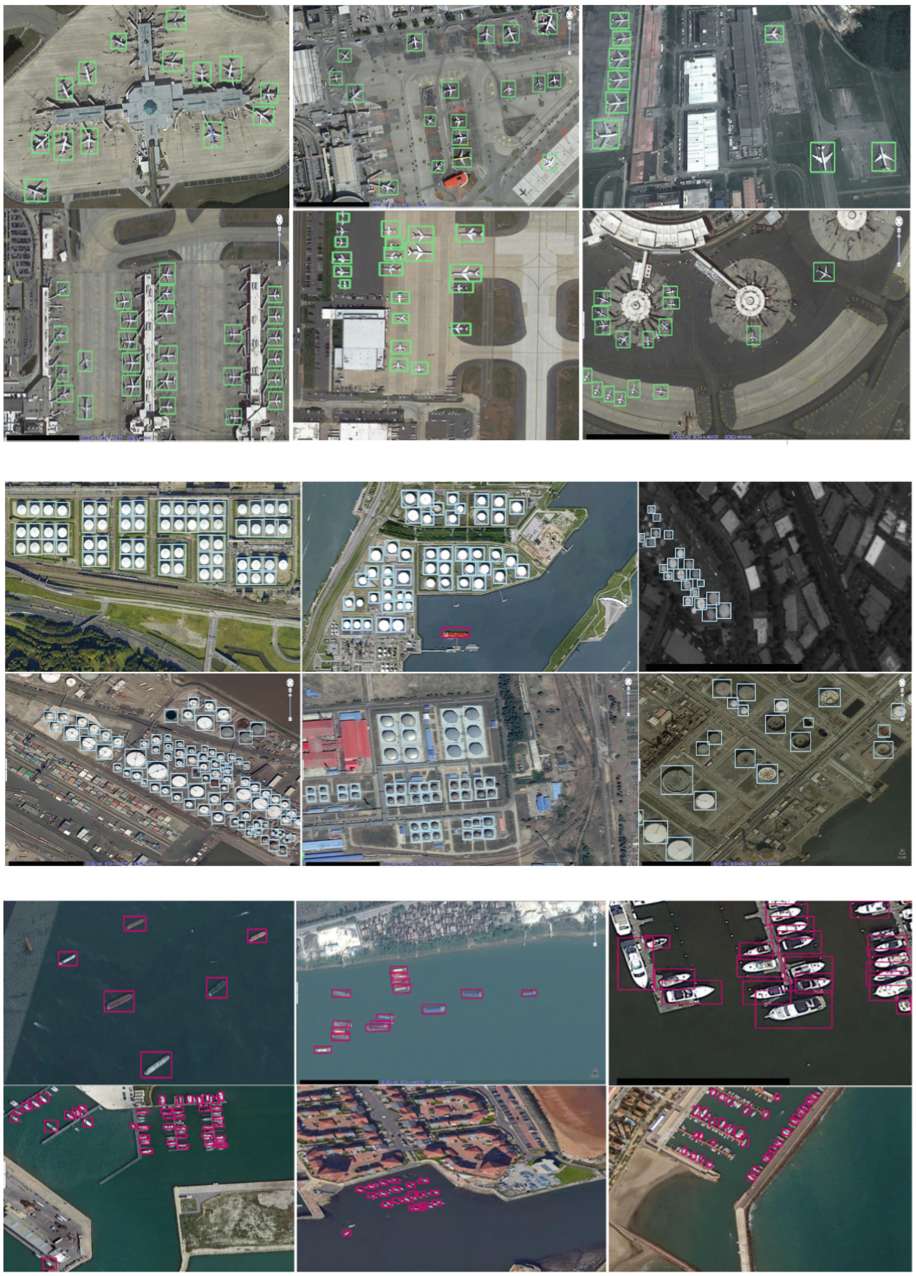

**Fig. 5.** Visual detection results on test dataset of our proposed S$^3$OD method. Small objects including airplane, oilcan, boat detection results are shown from top to bottom respectively. S$^3$OD has better performance in detection of dense and dispersed objects.

In order to illustrate the performance of the proposed train-from-scratch $S^3$OD in convergence, we train YOLOv3 in two ways including training without pretrained model and training with pretrained model to make a comparation. The Fig. 4 shows the training total loss value including x + y coordinates loss, w + h coordinates loss, confidence loss and class score loss. As Fig. 4 shown, green and blue curves present YOLOv3 without pretrained model and YOLOv3 fine-tuned with pretrained model respectively, red curve is our train-from-scratch S3OD with BatchNorm. Our trained $S^3$OD by BatchNorm from scratch has a better and stably convergence performance. And the mAP performance is better than two YOLOv3 models. These results indicate that using BatchNorm on each convolution layers is critical to train from scratch.

Figure 5 shows a few sample results from the RSSO test dataset and the corresponding detection is airplane, oilcan and boat which are small-densely or small-dispersedly distribution in remoting sensing images. The proposed $S^3$OD is capable of correctly detecting those small objects under various scenarios which are in low contrast, sparse texture and complex background. Besides, $S^3$OD is still prone to detection failure objects that are heavily overlapped with each other and will miss detecting objects which are too small to get the efficient feature. For this issue, we believe a better dilated convolutional network with a proper down-sampling factor and a better Non-Maximum-Suppression (NMS) can be adopted to address, which we will do in our feature work.

## 4 Conclusions

Aiming to improve the detection performance of small objects in remote sensing images, this paper presents an effective $S^3$OD method. A detection backbone with the small down-sampling factor is designed to keep high spatial resolution, two-branch dilated feature attention module is presented for small-dense and small-dispersed purposefully. Furthermore, BatchNorm is introduced to get a better training process for a robust detector. The experimental results on RSSO dataset demonstrate the effectiveness of the proposed method. Our proposed $S^3$OD pipeline exhibits strong competency in handling small object detection tasks. For future work, we will focus on the further tasks of small object detection and multi-scale object detection for remote sensing images.

**Acknowledgments.** The work was supported by National Natural Science Foundation of China (No. 91748131, No. 61771471, No. 61374159), the Youth Innovation Promotion Association Chinese Academy of Sciences (No. 2015112), the Foundation of CETC Key Laboratory of Data Link Technology (CLDL-20182316, CLDL-20182203), Natural Science Foundation of Shaanxi province (No. 2018MJ6048), and the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University (No. ZZ2019178).

## References


1. Wu, Y., Zhang, R., Li, Y.: The detection of built-up areas in high-resolution SAR images based on deep neural networks. In: Zhao, Y., Kong, X., Taubman, D. (eds.) ICIG 2017. LNCS, vol. 10668, pp. 646–655. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71598-8_57
2. Zhang, W., Wang, S., Thachan, S., Chen, J., Qian, Y.: Deconv R-CNN for small object detection on remote sensing images. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 2483–2486. IEEE, Valencia (2018)
3. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517–6525. IEEE, Honolulu (2017)
4. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
5. Ren, S., He, K., Grishick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS), Canada, pp. 91–99 (2015)
6. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 379–387 (2016)
7. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944. IEEE, Honolulu (2017)
8. Lin, T.Y., Goyal, P, Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007. IEEE, Venice (2017)
9. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv Preprint. arXiv:1804.02767 (2018)
10. Li, Y.H., Chen, Y.T., Wang, N.Y., Zhang, Z.X.: Scale-aware trident networks for object detection. arXiv Preprint. arXiv:1901.01892 (2019)
11. Pang, J., Li, C., Shi, J. Xu, Z., Feng, H.: R2-CNN: fast tiny object detection in large-scale remote sensing images. arXiv Preprint. arXiv:1902.06042 (2019)
12. Van Etten, A.: Satellite imagery multiscale rapid detection with windowed networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 735–743. IEEE, Hilton Waikoloa Village (2019)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv Preprint. arXiv:1409.1556 (2014)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for images recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV), pp. 770–778. IEEE, Amsterdam (2016)
15. Shen, Z., Liu, Z., Li, J., Jiang, Y.G., Chen, Y., Xue X.: DSOD: learning deeply supervised object detectors from scratch. In: IEEE International Conference on Computer Vision (ICCV), pp. 1937–1945. IEEE, Venice (2017)
16. Huang, G., Liu, Z.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. IEEE, Honolulu (2017)
17. Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., Sun, J.: DetNet: a backbone network for object detection. arXiv Preprint. arXiv:1804.06215 (2018)

18. Zhu, R., et al.: ScratchDet: training single-shot object detectors from scratch. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), accepted. IEEE, Long Beach (2019)
19. Santurkar, S., Tsipras. D., Ilyas, A., Madry, A.: How does batch normalization help optimization? In: Conference on Neural Information Processing Systems (NeurIPS), Montréal, pp. 2483–2493 (2018)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. IEEE, Utah (2018)
21. Cheng, G., Han, J., Zhou, P., Guo, L.: Multi-class geospatial object detection and geographic image classification based on collection of part detectors. ISPRS J. Photogrammetry Remote Sens. **98**(98), 119–132 (2014)
22. AIIA2018 Dataset Homepage. https://www.datafountain.cn/competitions/288/datasets
23. Everingham, M., Eslami, S.M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vision **111**(1), 98–136 (2015)