An Optimized Deep Neural Network Detecting Small and Narrow Rectangular Objects in Google Earth Images

Shenlu Jiang, Wei Yao, Man Sing Wong, Gen Li, *Student Member, IEEE*,
Zhonghua Hong, *Member, IEEE*, Tae-Yong Kuc, and Xiaohua Tong, *Senior Member, IEEE*

***Abstract*—Object detection is an important task for rapidly localizing target objects using high-resolution satellite imagery (HRSI). Although deep learning has been shown an efficient means of detection, object detection in HRSI remains problematic due to variations in object scale and size. In this article, we present a novel deep neural network (DNN) that combines double-shot neural network with misplaced localization strategy that adapts to object detection tasks in satellite images. This novel architecture optimizes the localization of small and narrow rectangular objects, which frequently appear in HRSI images, without accuracy loss on other size and width/height ratio objects. This method outperforms other state-of-art methods. We evaluated our proposed method on the NWPU VHR-10 public dataset and a new benchmark dataset (seven classes of small and narrow rectangular objects, SNRO-7). The NWPU VHR-10 dataset built a dataset for multiclass object detection; however, most labels are assigned in normal size and width/height ratios. SNRO-7 focuses on multiscale and multisize object detection and includes many small-size and narrow rectangular objects. We also evaluated the accuracy difference on DNN training and testing between gray scale and RGB datasets. The results of the experiment on object detection reveal that the mean average precision (MaP) of our method is 82.6% in NWPU VHR-10 and 79.3% in SNRO-7, which exceeds the MaPs of other state-of-the-art object detection neural networks. The model trained with the RGB dataset can achieve similar accuracy (around 79.0% MIoU) testing in both RGB and gray scale datasets. When training the model by mixing RGB and gray scale datasets in different ratios, the accuracy in the RGB channel significantly decreases with increasing gray scale images, but this does not influence the accuracy in the gray scale dataset.**

***Index Terms*—Artificial intelligence, object detection, optical image processing.**

I. Introduction

High-resolution satellite imagery (HRSI) has improved in terms of both its spatial and spectral resolutions. Therefore, the demand for rapidly localizing targets has increased for applications, such as aircraft detection [1]–[3], ship detection [4]–[6], and vehicle detection [7]–[10], especially Li *et al*. [9] revealed a rotatable region-based deep neural network (DNN) for vehicle detection, which performs well on oriented vehicle detection in the aerial image. Object detection is a basic task in HRSI analysis. Two classes of methods are used for traditional target detection: 1) dividing the objects into patches based on group relationships among the pixels, and 2) scanning scenes based on classifiers from manually labeled features. According to previous reports [11]–[13], the first strategy is sensitive to the complicated backgrounds in HRSI. Methods based on a scanning window are affected by the quality of both the human-crafted features and the training data. These limitations restrict the application of the two classes of detection methods.

Convolutional neural networks (CNNs) use a detector to directly learn the objects' features instead of using human designated features. Existing object detection approaches based on deep learning have achieved exceptional object detection accuracy [14]. CNNs self-learn the features through hidden layer descriptions that provide features that are more appropriate for machine use. Object detection neural network localizes the rough position and classifies the objects in the images. Comparing with instance segmentation, the object detection neural network generally occupies quicker speed. In this article, the training architecture and dataset construction are based on the rule of object detection. The training and inference times of object detection DNNs are fast, and the dataset on labels is easily acquired. The semantic/instance segmentation is a pixel-level classification problem that categories each pixel on the image and have been applied on remote sensing task, such as lane segmentation by using optimized FCNDensenet [15]. Otherwise, optimizing the inline relationship through spatial relation

Manuscript received August 2, 2019; revised November 16, 2019, December 27, 2019, and January 31, 2020; accepted February 15, 2020. Date of publication March 9, 2020; date of current version April 2, 2020. This work was supported in part by the National Key R&D Program of China under Grant 2018YFB0505400, in part by the National Natural Science Foundation of China under Grant 41871325, in part by the Korea Evaluation Institute of Industrial Technology (KEIT) funded by the Ministry of Trade, Industry & Energy (MOTIE) under Grant 1415162366, Grant 1415162820, and Grant 1415164140, in part by Shanghai Rising-Star Program under Grant 17QB1403800, and in part by a grant PolyU 1-BBWD from the Research Institute for Sustainable Urban Development, the Hong Kong Polytechnic University. *(Corresponding author: Zhonghua Hong.)*

Shenlu Jiang, Gen Li, and Tae-Yong Kuc are with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon 440-746, South Korea (e-mail: sljiang@skku.edu; ligen@skku.edu; tykuc@skku.edu).

Wei Yao and Man Sing Wong are with the Department of Land Surveying and Geo-Informatics, the Hongkong Polytechnic University, Hong Kong (e-mail: wei.hn.yao@polyu.edu.hk; ls.charles@polyu.edu.hk).

Zhonghua Hong is with the College of Information Technology, Shanghai Ocean University, Shanghai 201306, China, and also with the Shanghai Tuyao Information Science and Technology Co., Ltd, Shanghai 201306, China (e-mail: hzh_2000@163.com).

Xiaohua Tong is with the College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China (e-mail: xhtong@tongji.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.2975606

module for scene understanding in aerial view segmentation was revealed in [16]. But semantic segmentation is extremely time-consuming to label the dataset and train the DNN. As the aim of this study was object detection, we compared the proposed DNN with the state-of-the-art object detection DNNs. The methods related to the segmentation will be discussed in our future papers.

Two major trends for object detection based on deep learning neural networks have developed: two-stage and one-stage detection. Two-stage detection is inspired by the traditional sliding window strategy, which uses a region of proposal step to search for candidates and then inputs the candidate regions into a neural network to obtain a score for each object [17]–[19]. One-stage detection considers detection of a regression task, which searches for objects using the neural networks and feature maps [20], [21]

Approaches based on CNNs achieve exceptional accuracy in object detection [14], much better than traditional detection methods. The following four-factor challenges remain for object detection in HRSI.

1) The targets to be detected in HRSI have different scales.
2) The limits of HRSI sizes blur small-scale objects, complicating detection.
3) The shape of some targets in remote sensing imagery is narrow rectangle, preventing accurate localization.
4) One-channel gray scale images (such as panchromatic images) and multichannel images (such as multispectral images) are widely employed in HRSI.

According to these four issues, a successful detector should have the strength to overcome the challenges of scale, image quality, overlapping objects, and detection, in both three-channel RGB and one-channel gray scale HRSI images.

Regarding the state-of-the-art deep learning methods in the HRSI field, the sliding window plus image classification strategy [22] is commonly used, but the millions of steps in object detection restrict HRSI application. Estimating potential regions through feature maps [10] is also an option, but small-scale object proposals and the localization of blurry objects remain problematic. The risk of overlapping may restrict the performance of traditional methods. However, DNNs can ideally classify different objects in overlapping regions. This has been demonstrated with the PASCAL visual object classes (VOC) 2012 semantic segmentation benchmark [23]. Many reports have been published on target detection using multispectral images. However, research work is limited on target detection using one-channel gray scale images [24]. In this work, we compared and analyzed the performance loss between detection in three-channel RGB images and one-channel gray scale images. Various DNNs demonstrate that inline connection among the objects is the key to accurately grouping and classifying the objects.

To address the abovementioned challenges, a novel neural network was employed to maintain robust and accurate detection in HRSI. Our neural network uses basenet to transfer the features by learning the visual geometry group (VGG16) [25]. To adapt the aerial view, an extended version of VGG [26] were reported trained by using extra aerial view datasets, such as NWPU-RESISC45 [27]. A multilabel classification DNN is also revealed in [28] to adapt aerial view scene understanding. The improved basenet and dataset allow the DNN to adapt the features to aerial view. We then combined the double-shot neural network with the misplaced localization strategy to detect multiple objects in Google Earth satellite images. This combined approach provides a novel architecture to improve the detection of small/blurred objects. The proposed method can also be used for multiple object detection.

The rest of this article is organized as follows. Section II addresses the architecture of the components of our neural network for multiple-class object detection. Section III introduces the experimental datasets extracted from Google Earth satellite imagery and outlines the comprehensive experiment we conducted to evaluate the performance of the proposed method for both three-channel RGB images and one-channel gray scale images. The results are discussed in Section III. Finally, Section IV concludes this article.

II. Proposed Method

Faced with challenges in object detection in satellite imagery, we propose a novel neural network combining double shot with misplaced strategy to optimize the detection for small/blurred and narrow rectangle objects.

To improve detection capability in the multiscale context, the neural network employs a series of 3 × 3 convolutional layers to decrease the size of the feature map to describe features on different scales. Our approach employs the double-shot strategy combined with anchor box technology to enhance detection capability for small/blurry and narrow rectangular objects. After each convolutional layer, a convolutional 1 × 1 extra layer with a misplaced size is used to increase object representation capability. Unlike other state-of-the-art methods, our neural network extracts the feature map twice with different remapping sizes. This provides a more efficient regional proposal step and better prediction ability for objects (especially in cases where skies are unclear or target objects are small). Finally, all the convolutional layers output the multibox layer and then output the results of object detection using SoftMax. To filter the overlapped proposals, nonmaximum suppression (NMS) is used to select the most ideal region for the target object.

Fig. 1 presents the architecture of the components of our neural network. It consists of three steps: feature extraction from Basenet, extra feature extraction, and object detection and NMS. The details of all the components are discussed in the following sections.

A. Basenet

Initially training a new neural network is a difficult task because of the required initialization of the original parameters of the nodes inside the neural network. To solve these issues, fine-tuning is commonly used in DNNs to achieve sufficient training quality with a smaller number of training iterations. Fine-tuning is achieved by employing a pretrained neural network to extract the features and adjust the parameters inside the feature map layers. This allows the neural network to have an original parameter set better suited for the target dataset.

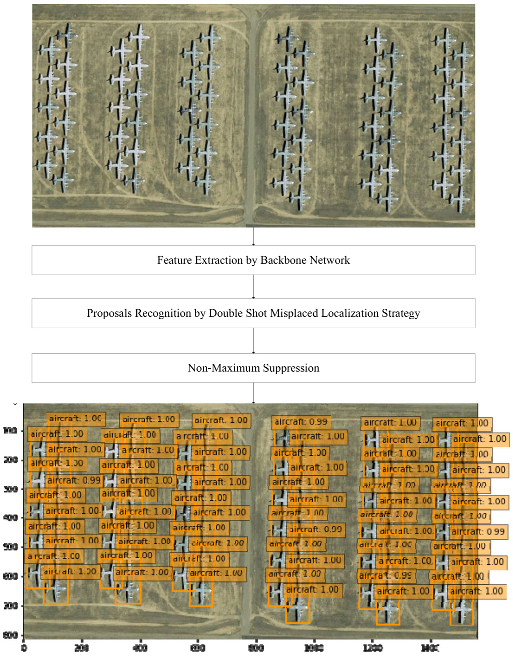
Feature Extraction by Backbone Network
Proposals Recognition by Double Shot Misplaced Localization Strategy
Non-Maximum Suppression

Fig. 1. Architecture of our proposed DNN.

In our neural network, a VGG and ResNet pretrained by ImageNet [29] is employed to fine-tune our model to the dataset. Since the proposed neural network employs a large input size (600 × 600 pixels) image to extract more detailed features and describe small objects, the VGG has the most advanced feature-extracting capabilities. However, its computational costs are higher than other back nets, such as ZFNet [30]. VGG has lower computing costs, but just slightly worse feature extraction than the residual network (Resnet) [31]. However, ResNet performs better in feature extraction. In current DNNs, the ResNet and VGG are widely used as backbones for the task of object detection. Therefore, we employed both VGG and ResNet as the basenet to test the accuracy of our proposed method. In all detection steps, the images are first input to the basenet to extract the image features. When using VGG, all the layers before conv4_3 are set to be unchanged to guarantee feature extraction strength from the pretraining procedure. When using ResNet, our proposed method directly utilizes the Pooling 4 as the inputting features. In the extra layers, the residual convolutional layer is employed when using the Resnet.

B. Extra Feature Layers

After the feature maps are extracted by the basenet, our neural network employs a series of convolutional layers to extract the features of objects in the input HRSI. This is conducted to allow the neural network to understand the image.

As shown in Fig. 2, our neural network used the double-shot strategy to extract extra feature layers. After the basenet, the neural network employs a 3 × 3 convolutional layer (stride 1 and padding 2) to downsize the size of the layers from 19 × 19 to 1 × 1. Our neural network employs a 1 × 1 convolutional layer (stride 0 and padding 1) for every convolutional layer to produce double shots. The 3 × 3 convolutional layers with stride 1 and padding 2 ensures that the neural network can gradually extract features on multiple scales. The other methods [single-shot multibox detector (SSD), you only look once (Yolo), RetinaNet, and fast region-based CNN (Faster RCNN)] generally employ a top-down strategy to extract the features from large scale to small scale by gradual downsampling. However, these neural networks are generally used to detect the general view objects. The objects in the proposals without the misplaced strategy are used by SSD and Faster RCNN. However, remote sensing imagery is generally only a small proportion, and the feature context is different from that of the general view objects. Therefore, adapting the features to the remote sensing objects is essential for the neural network. To solve this problem, the double-shot strategy is employed to improve small object description. This strategy provides a random prediction of the objects' appearance with a 1 × 1 convolutional layer to address issues, such as blurring or partial covering. In addition to improving feature extraction, the double-shot strategy also collaborates with the misplaced strategy (see Section II-D) to better suit the detection needs of the remote sensing community. In the extra feature layers, the double-shot strategy ensures the strength of multiscale feature extraction and provides a sufficient description of the objects.

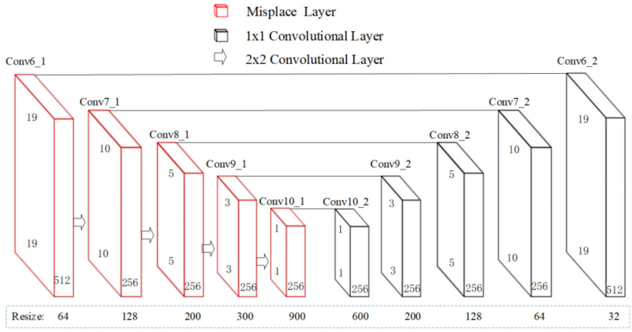

Fig. 2. Architecture of extra layers of our proposed DNN.

C. Detection

After the features are extracted on multiple scales, the detection layer is used to estimate the potential positions of the objects. In our neural network, we designed this step as a regression task and employed default boxes to estimate the object proposals. To output the exact position of the objects in the scene, each proposed box is defined by four parameters: its center (*x, y*), weight, and height. For each feature map layer (conv4–conv10), every cell of the feature map is used as a default feature map that provides the *x* and *y* coordinates of the object on the scene. It is then multiplied by the scale factor

$$P_{x,y,w,\ h} = [\text{Random }\ (x, y)\ , ar \times (w, h)] \tag{1}$$

where *x* and *y* are the coordinates on the image; *w* and *h* denote the proposals' weight and height, respectively; and the proposal region *P* in each extra feature is defined by the random position proposal (*x, y*) on the image, which multiplies the weight w and height h by the anchor box ratio ar.

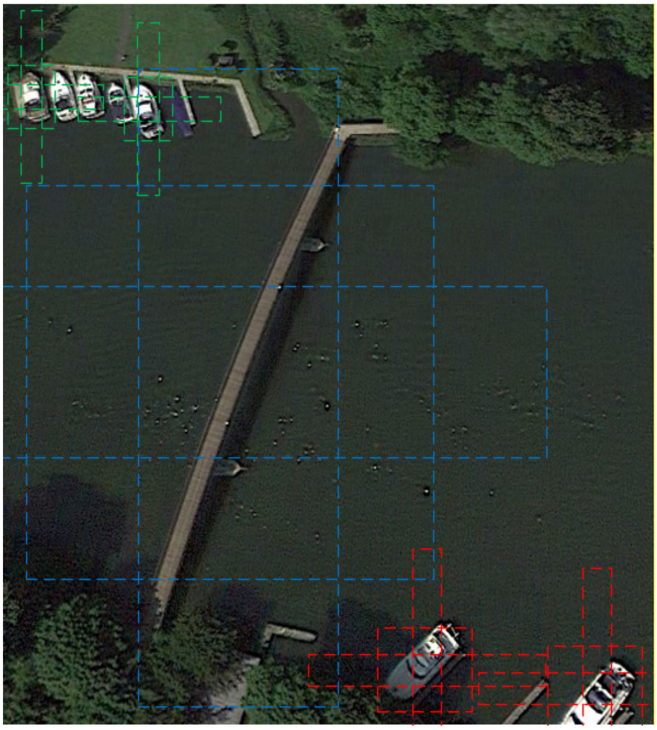

Fig. 3. Visualization of localization results on different scales (red, blue, and green present different scales).

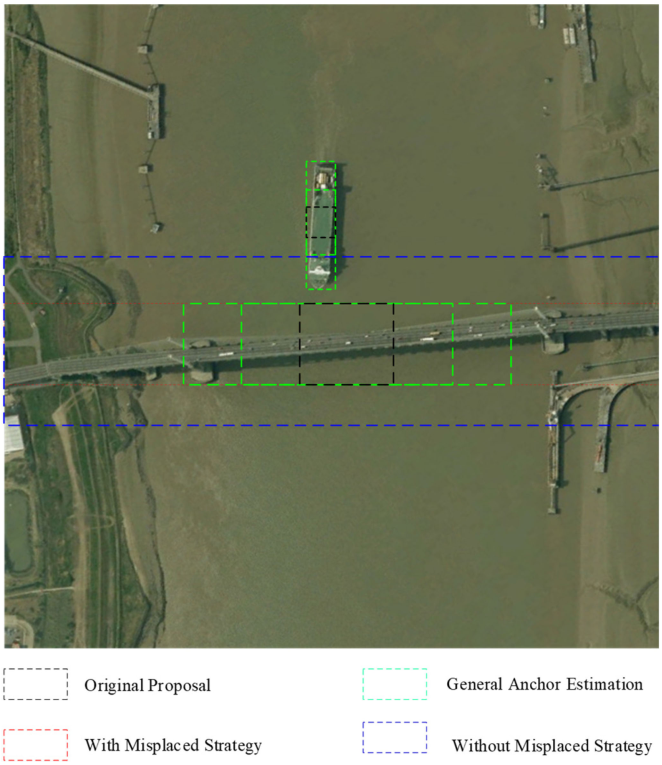

Fig. 4. Comparison of the visualization quality both with and without the proposed misplaced localization strategy.

D. Anchor Box

Each cell of the feature map only has a 1:1 ratio. This cannot accurately describe an object's shape in HRSI. Therefore, the anchor box is a technology used to expand the proposals during detection, which has been widely used in state-of-the-art DNNs for object detection.

The other methods, e.g., SSD and Faster-RCNN, estimate the proposals, such as the box without using the misplaced strategy. This is not sufficiently precise for localization in remote sensing objects. As shown in Fig. 3, the anchor box adds extra boxes with different ratios during the estimation of an object's potential region. This design can expand the coverage field and reduce the number of estimations. The ratios of the anchors are generally 1:2 or 1:3.

However, this design cannot be used to accurately estimate the proposals for the objects in HRSI, such as the bridge shown in Fig. 4. The width of the bridge occupies the entire image, but its height is narrow. In this case, the original size of the default box should be large enough for the 1:3 ratio to cover the entire bridge (blue box). However, this is not a good representation of the bridge.

E. Combining Double Shot With Misplaced Localization Strategy

To solve this problem, we present a misplaced localization strategy for the default box's size combined with the double-shot strategy. Notably, since too many parameters are included in the convolutional layer, simply adding the ratio for the default box in each convolutional layer would increase the convergence difficulty of the neural network. Thus, our neural network uses the misplaced size of the anchor for the double-shot layers to provide better object-locating performance.

To detect the feature maps on different scales, a remapping step is essential for representing the objects in their original size. This is accomplished by multiplying the scale value by the feature map size (e.g., a 19×19 layer is multiplied by 32 to represent positions in the 600×600 image). As shown in Fig. 2, for the first shot, each layer remaps the object to the original image using its previous layer's size, whereas the second shot remaps by the original size. conv_9 is multiplied by 300 and conv_10 is multiplied by 900, not their previous layers' sizes, because the final outputs are much larger than the size of the original image. Based on experience, the width/height ratio for different classifications vary, i.e., a bridge may cover ratios from 1:3 to 1:8, whereas a ship or aircraft may only range from 1:1 to 1:2.

Objects with a high width/height ratio usually occupy a large region, and a lower ratio usually indicates objects are smaller. Due to the second shot having stronger feature extraction, we used the second shot to localize general objects with the normal anchor size. The first shot uses the misplaced strategy to localize distorted objects. Considering the shape of the objects in HRSI, from our experience, we set ratios of 1:1, 1:2, and 1:3 to the layers with large sizes (conv6_1 to conv7_1 and conv6_2 to conv7_2) to ensure that the regions rich in small-scale detail could be used to rapidly locate objects, such as ships, aircraft, playgrounds, or storage tanks. The ratios for small-scale layers (conv8_1 to conv10_1 and conv8_2 to conv10_2) were set to 1:1 and 1:2 to enable detection of large objects, such as bridges, ports, or viaducts.

Since the misplaced strategy has already described objects in the distorted region and the second shot for small-size layers, it can also detect their approximate sizes. Notably, the conv4_3 and FC6 layers only employ the double-shot strategy, and their first shots are not used as outputs into the SoftMax layer because these two layers have large sizes that are used to detect small objects. Small objects are less distorted than large-scale objects. Therefore, only the double-shot strategy is employed to improve feature extraction and assist in the localization of small objects

$$O^n_{x,y,w,\ h} = \begin{cases} P^n_{x,y,w,\ h} \times s_{k-1} & \text{if } s \text{ is the first shot} \\ P^n_{x,y,w,\ h} \times s_k & \text{if } s \text{ is the second shot.} \end{cases} \quad (2)$$

The output region $O^n_{x,y,w,\ h}$ on different scales (n) is defined by the proposed $P^n_{x,y,w,\ h}$ multiscale ratio s_k; the scale ratio is determined by whether it is the first or the second shot. After processing with the neural network, each detection layer outputs the objects' potential proposals. They are then input into the SoftMax layer to compute the probability and classify each target.

F. Anchor Box NMS

After all objects' proposals are estimated, our program determines the 2000 estimated boxes with the highest confidence rate for the next step. The threshold is set by the number of proposals whose confidence rate is higher than 0.01. As shown in Fig. 1, 1586 proposals have a confidence rate higher than 0.01. The threshold is set to 2000 for two reasons. First, the NMS employs brute force strategy, which is increasingly time consuming with increasing number of proposals. Then, the number 2000 can ensure both efficiency and accuracy [32]. SSD and Faster RCNN also employ 2000 as the threshold. As shown in Fig. 5, the box with the highest confidence rate is chosen from overlapping boxes if the intersection over union (IOU) is >0.7. This filters the proposals in larger or smaller regions out of the ideal proposal

$$R = \frac{\left|O^n_{x,y,w,h} \cap O^m_{x,y,w,h}\right|}{\left|O^n_{x,y,w,h} \cup O^m_{x,y,w,h}\right|}. \quad (3)$$

According to (3), each final output result R must meet the constraint that no output proposal $O^n_{x,y,w,\ h}$ overlaps with $O^m_{x,y,w,\ h}$ larger than an IOU of 70%. After NMS, our object detection neural network is finished, and the ideal results of objects' positions are output. To avoid the issue caused by of overlapping, we only operate NMS for the same classification. This avoids the filtering of other targets during the procedure.

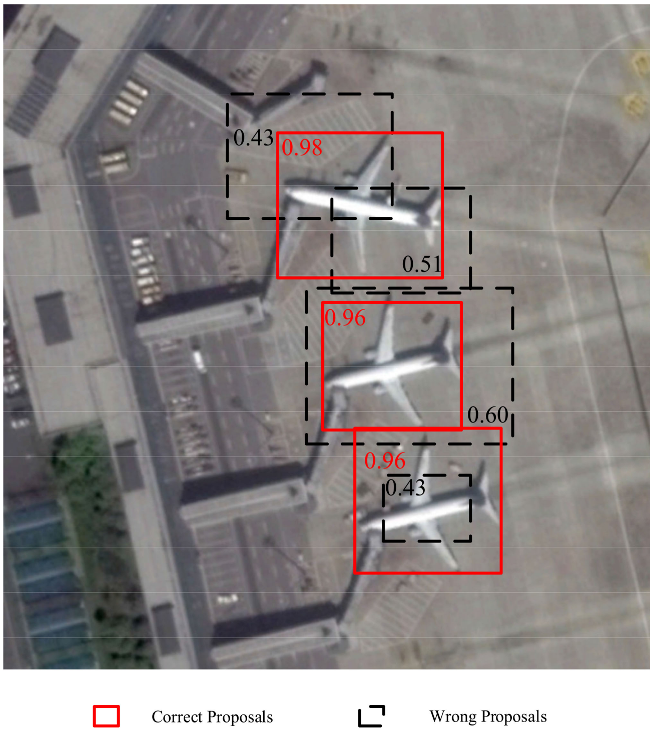

Fig. 5. Visualization of NMS.

G. Optimizing the Proposed Approach

In terms of the ground view, our DNN focuses more on the accurate localization of small objects and the output of more accurate regions of interest (ROIs) than other state-of-the-art methods. The width/height ratios of the objects on the ground view are usually lower than those in the remote sensing community. Most ground-based neural networks regard the results as correct detection when the intersection region between the benchmark and the output is over 50%. The evaluation criteria can be used in the ground view, but they are different from the requirements in the remote sensing community. The other factor is small object detection, to which most state-of-the-art methods pay little attention because small objects only comprise a tiny percentage of common benchmarks of the computer vision community, i.e., PASCAL VOC, and Microsoft common objects in context (MS COCO) [33]. Thus, a limited number of optimizations are used for this part. Our DNN focuses on these drawbacks and provides a new strategy to fix the abovementioned issues, meeting the requirements for object detection in the HRSI community.

The major difference between other state-of-the-art methods (most single-stage and two-stage methods) and the proposed neural network is shown in Fig. 6. Regarding current state-of-the-art methods, almost all neural networks estimate the proposals of the objects on feature maps through a single path, i.e., extract the extra-feature maps from large to shallow and localize all objects in different sizes and width/height ratios through the same chain. Unlike object detection in the ground view images, many objects are small and narrow rectangular in shape from the aerial view. The neural network struggles to obtain the same region proposals as the ground view images because one feature map cannot describe the objects of all sizes and proposal anchor may inevitably lead to proposal conflict if blindly adding the ratios. To solve the abovementioned challenges, our neural network applies a novel strategy that divides the network into two paths to split objects in different features

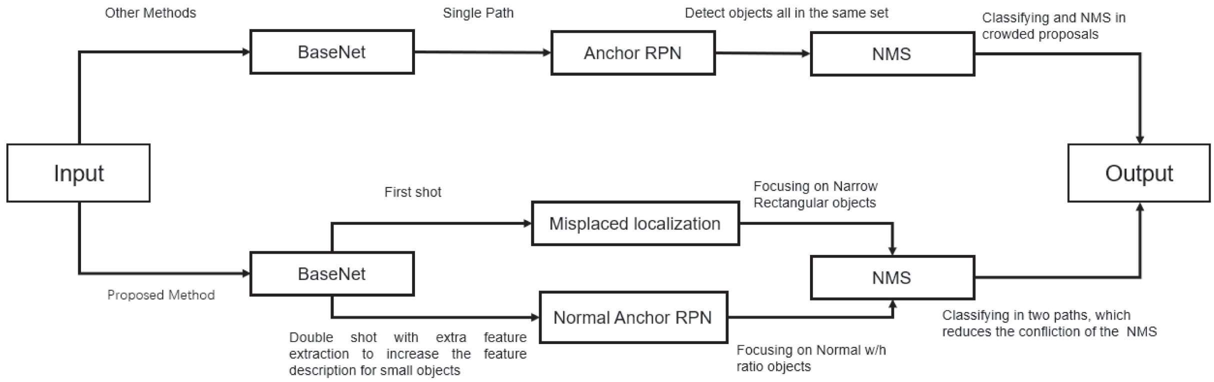
Other Methods
BaseNet
Single Path
Anchor RPN
Detect objects all in the same set
NMS
Classifying and NMS in crowded proposals
Input
Output
First shot
Misplaced localization
Focusing on Narrow Rectangular objects
BaseNet
Proposed Method
NMS
Normal Anchor RPN
Classifying in two paths, which reduces the confliction of the NMS
Double shot with extra feature extraction to increase the feature description for small objects
Focusing on Normal w/h ratio objects

Fig. 6. Difference between existing methods with our proposed DNN.

to be localized in different paths. The double shot combining misplaced localization strategy enables the detection of different types of objects. The first shot collaborating with misplaced strategy focuses on detecting narrow rectangular objects, which are hard to accurately localize using the normal anchor ratio. The second shot with double-shot strategy focuses on detecting the normal width/height ratio objects on different scales. As the ratio in anchor is limited and the feature map is further extracted, the second shot has the ability to detect very small-sized objects (small-size objects generally detected by the 1:1 ratio anchor).

As shown in Fig. 7, the main idea of our combined approach is to divide two branches for different size ratio objects to avoid conflict during proposal estimation and NMS. The first shot focuses on detecting the narrow rectangle objects, and the second shot focuses on detecting smaller objects. Combining the two strategies, sufficient and well-extracted feature proposals allow the neural network to detect both rectangular objects and small/blurry remote sensing objects in HRSI imagery.

This is not simple optimization, but an innovative architecture. The architecture allows the localization to avoid conflicts of objects in different scales and width/height ratios, which provides a more reasonable architecture for localizing objects with different scales and sizes. With this strategy, the proposed architecture can considerably improve detection accuracy as demonstrated by the experimental results.

III. Experiments

We used experiments to evaluate the proposed neural network. This section introduces the platform and training procedures, provides an evaluation of the neural network for the public NWPU VHR-10 dataset [34], outlines the details of the collected dataset, describes the evaluation of the accuracy of our neural network and the comparison with other state-of-the-art methods in our dataset and in other public datasets, and describes the assessment of the performance of object detection using our neural network for both three-channel RGB images and one-channel gray scale images.

A. Platform and Training Procedure

Our neural network was deployed using Caffe on the Ubuntu operating system 16.04. This provides good compatibility with

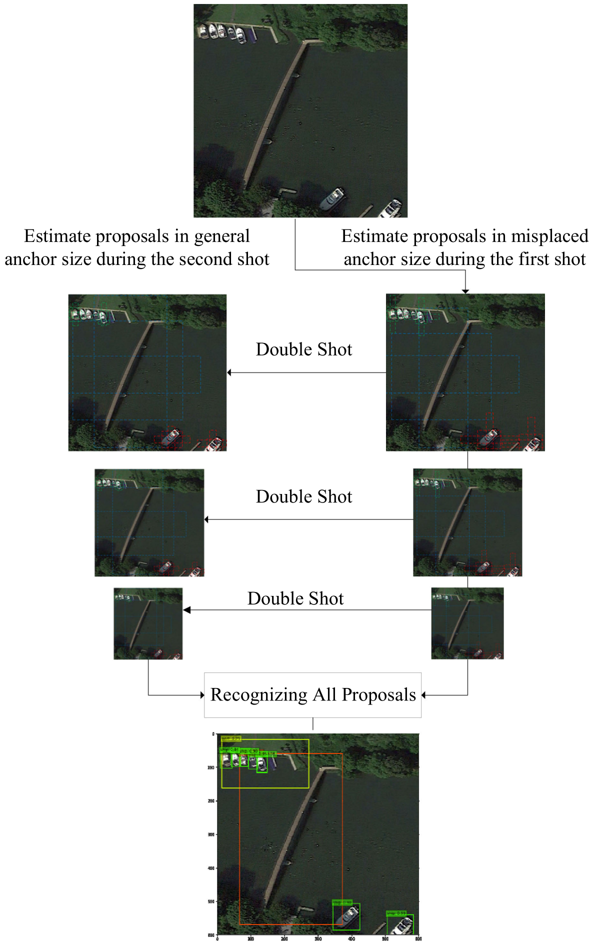
Estimate proposals in general anchor size during the second shot
Estimate proposals in misplaced anchor size during the first shot
Double Shot
Double Shot
Double Shot
Recognizing All Proposals

Fig. 7. Process of combined misplaced localization strategy and double-shot neural network.

the graphics processing unit (GPU) platforms. Our desktop uses an I7 7700 central processing unit with 16 GB of memory and a GTX 1080 GPU. This allows high-performance computation for deep learning. To maximize the GPU memory usage and rapidly reach the convergence point, we set the batch size to 8 (7819 MB in 8 GB). We set the learning rate to 0.005 with a learning rate decay of 1/10 for every 60 000 iterations. The dataset had a total of 250 000 interactions and SGD Momentum

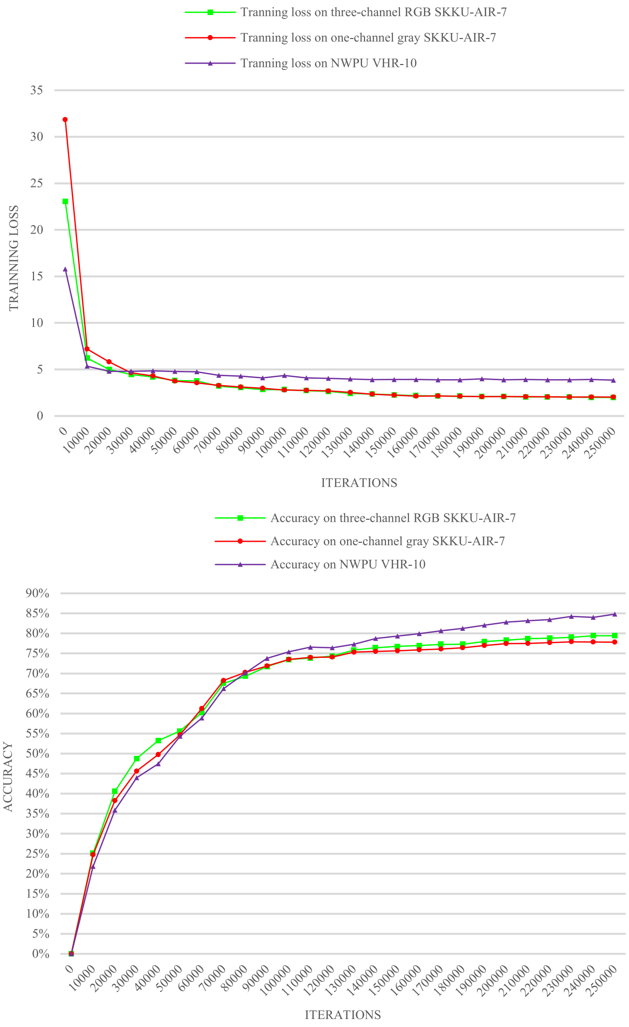

Fig. 8. Training procedures of RGB and gray scale datasets.

optimizers. There were 6000 decay steps and 25 000 iterations in NWPU VHR-10. We selected a large total training step number to ensure sufficient and complete training of the neural network to maximize its performance. The multistep number and decay rate were set according to our experience. The loss rate of our neural network stabilized after 60 000 iterations. Additional training steps required considerable computing time, but with limited mitigation of loss. The weight decay rate was set to 0.1 because the neural network is stable under the training of an appropriate learning rate, and the loss is mitigated only when the learning rate is significantly lower. According to Rude [35], different optimizers only have limited accuracy offset after neural networks converge.

The training procedures of the proposed method in the NWPU VHR-10 and that in SNRO-7[1] (seven classes if small and narrow rectangular objects, SNRO-7) [36] for both three-channel RGB images and one-channel gray scale images are shown in Fig. 8. The loss rates for our neural network for both three-channel RGB and one-channel gray scale images drop to about 4.0 at step 1 (60 000 interactions); they gradually decrease to about 2.0 by the end of training. Although the three-channel RGB images are equivalent to those in the one-channel gray scale dataset, the trends of both the three-channel RGB images and the one-channel gray scale images are similar. This illustrates that destroying the three-channel RGB images does not influence the coverage of the neural networks.

B. NWPU VHR-10 Quantitative Evaluation

The NWPU VHR-10 contains 10 classes of geospatial objects, with a total of 650 labeled images. Because the dataset was not divided into training and test datasets, we randomly selected 450 of its images as the training dataset; the remainder was used as the test dataset.

To evaluate our proposed method, we compared our neural network with three other state-of-art methods. In the comparison tests, the SSD [21], Faster RCNN [19], and Yolo [20] were selected due to their generally accepted advances in object detection. We used the mean average precision (MaP) to represent their accuracy, which is calculated as

$$\mathrm{MaP} = \frac{\sum c}{N} \tag{4}$$

where the precision rate of each class is defined as c and the number of classes is defined as N. The MaP can evaluate the average performance of the model for detecting different objects. IOU values greater than 0.7 between the results and test dataset were marked as true positive results.

The results presented in Table I illustrate that our proposed method was the most accurate with an 84.8% MaP on average. Double shot combined with the misplaced strategy assisted our proposed method in the detection of small items, such as airplanes, ships, and land vehicles, and allowed the neural network to localize objects with distorted shapes (e.g., bridges) with higher accuracy than the other methods. For the other general objects, the accuracy of our proposed method was similar to that of the others. However, due to the limited number of objects and images contained in the NWPU VHR-10 dataset, the neural network may have suffered from overfitting. Thus, the dataset does not highlight the overall performance of the neural networks. Some of the labels were missed in the images, as shown in Fig. 9. Examples of this include a harbor with unlabeled ships, mislabeled storage tanks near the boundary of an image, or mislabeled vehicles on golf courses. This distorts both coverage during training and accuracy during testing. Apart from our experiments, the baseline of the NWPU VHR-10 dataset was reported by Cheng *et al.* [34]. The accuracies of Faster RCNN and other methods are relatively lower than reported previously [34] for three reasons.

1) Cheng *et al.* [34] extended the dataset but we did not apply data augmentation in this article.
2) Cheng *et al.* [34] employed 50% overlapping region as the true positive but ours was 70%.
3) Cheng *et al.* [34] used Titan X as the GPU platform, which has a higher batch size and computing speed comparing with ours (GTX 1080).

[1] SNRO-7 is available: http://dx.doi.org/10.21227/j7nx-2495.

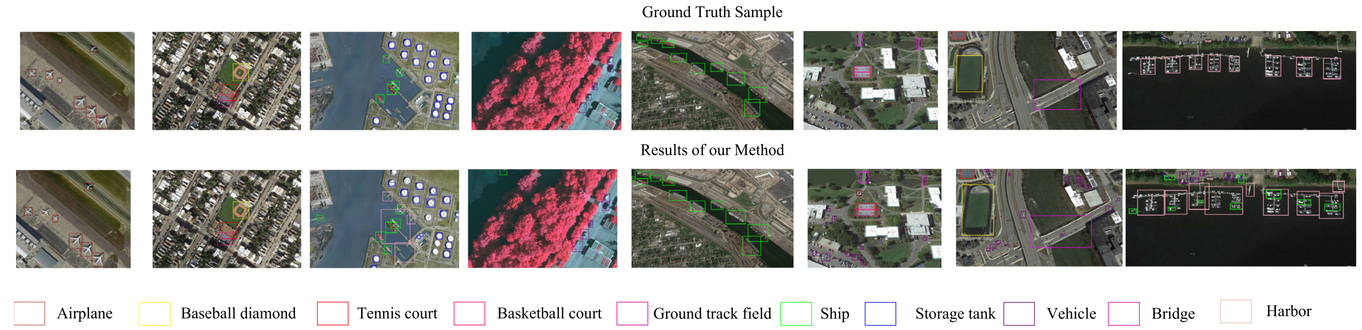

Fig. 9. Results of our proposed method in NWPU-VHR-10 and its ground truth.

TABLE I
MAP OF THE OBJECT DETECTION METHODS ON THE NWPU VHR-10 DATASET

TRAINED BY RESNET

Item	SSD	YoloV3	Faster RCNN	Retina Net	Ours
Airplane	0.908	0.806	0.925	0.912	0.930
Ship	0.791	0.825	0.830	0.828	0.845
Storage tank	0.848	0.784	0.810	0.885	0.871
Baseball diamond	0.913	0.879	0.898	0.938	0.928
Tennis court	0.790	0.780	0.850	0.830	0.820
Basketball court	0.878	0.753	0.820	0.859	0.890
Ground track field	0.580	0.635	0.930	0.794	0.780
Harbor	0.620	0.601	0.793	0.735	0.760
Bridge	0.500	0.543	0.641	0.788	0.810
Vehicle	0.810	0.790	0.750	0.860	0.845
Mean AP	0.764	0.740	0.825	0.843	0.848

TRAINED BY VGG

Item	SSD	Faster RCNN	Ours
Airplane	0.904	0.920	0.924
Ship	0.802	0.843	0.793
Storage tank	0.865	0.778	0.871
Baseball diamond	0.930	0.892	0.932
Tennis court	0.795	0.835	0.810
Basketball court	0.871	0.810	0.893
Ground track field	0.575	0.920	0.758
Harbor	0.623	0.783	0.725
Bridge	0.500	0.652	0.728
Vehicle	0.803	0.732	0.830
Mean AP	0.766	0.816	0.826

FASTER RCNN IN DIFFERENT ANCHOR SIZE AND RATIO

Item	Ground	Aerial	Max
Airplane	0.913	0.920	0.895
Ship	0.835	0.843	0.852
Storage tank	0.781	0.778	0.770
Baseball diamond	0.887	0.892	0.905
Tennis court	0.809	0.835	0.830
Basketball court	0.815	0.810	0.820
Ground track field	0.920	0.920	0.916
Harbor	0.780	0.783	0.768
Bridge	0.659	0.652	0.670
Vehicle	0.743	0.732	0.703
Mean AP	0.814	0.816	0.812

Therefore, the accuracy distribution shown in Table I is slightly lower than that in [34]. However, our proposed DNN is more accurate in the same platform and evaluation index compared with state-of-the-art methods, which could prove the success of our proposed DNN.

The NWPU VHR-10 dataset contains airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. The dataset contains multiclassification, which is suitable for multiclass object detection. As the dataset is collected on different scales, some targets, i.e., airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, and harbor, are large-sized objects. As the faster RCNN were used in many papers as the baseline, we further compare the results with various anchor sizes and ratios to evaluate influence of the proposal anchor. There are three sets of experiments, the ground reveals original ground view set (size $\{128,\ 256,\ 512\}$ and ratio $\{$0.5:1, 1:0.5, 1:0.5, 1:1 1:2, 2:1$\}$), the aerial means the parameters optimized aerial view set (size $\{64,\ 128,\ 256,\ 512\}$ and ratio $\{$0.5:1, 1:0.5, 1:1, 1:1.5, 1.5:1, 1:2, 2:1$\}$), and the max means the maximized number of size and ratio set (size$\{32,\ 64,\ 128,\ 256,\ 512\}$ and ratio $\{$0.5:1, 1:0.5, 1:1, 1:1.5, 1.5:1, 1:2, 2:1, 1:3, 3:1$\}$). In the results, we can find out that the median number of size and ratio obtained the best results. The results illustrate that blindly increasing the anchor size and ratio leads to the difficulty on training coverage, single path feature maps cannot cover sizes objects, and the parameters for ground view set is not suitable for the aerial view image objects, which proves the essential benefit of the double shot with misplaced localization strategy that opens a new path to localize different sizes and ratios objects in specific feature divisions. However, due to the limited number of objects and images contained in the NWPU VHR-10 dataset, the neural network may suffer from overfitting. Thus, it is hard to demonstrate the complete performance of the neural networks. Some of the labels were missed in the images, as shown in Fig. 9. Examples of this include a harbor with unlabeled ships, mislabeled storage tanks near the boundary of an image, or mislabeled vehicles on a golf course. Few narrow rectangular objects were included, and the number of small objects was not sufficient, which is why we built the new dataset SNRO-7.

C. SNRO-7 Details

To further evaluate our neural network, we built up a new dataset (SNRO-7). Some of images collected from an AID image recognition dataset [37] and others are from Google Earth. The AID dataset collected from Google Earth history images contained 30 classes. However, some of the classifications, e.g., beach, bare land, and commercial regions, were ill-suited for object detection. Thus, we selected scenes with ships, aircraft

TABLE II
PROPERTIES OF OBJECTS IN THE DATASET

Item	No.	Min width/height	Max width/height
Ship	20,865	10 × 10	80 × 80
Aircraft	4050	15 × 15	100 × 100
Playground	446	20 × 20	120 × 120
Viaduct	422	200 × 200	800 × 800
Port	1444	300 × 300	700 × 700
Storage Tank	6019	10 × 10	100 × 100
Bridge	900	120 × 120	1000 × 1000

TABLE III
DISTRIBUTION OF RESOLUTION FOR OUR DATASET

Image Size (Numbers)	**Image Resolution (Numbers)**
≤600 × 600 (1493)	<0.5 m (1464)
600 × 600–000 × 1000 (121)	0.5–1 m (841)
>1000 × 1000 (799)	>1 m (108)

(from airports), playground, viaducts, ports, storage tanks, or bridges. Since the AID is an image recognition dataset, we labeled the selected images manually. The size of the AID dataset was 600 × 600, but the size of the remote sensing imagery usually exceeded this ratio. Therefore, we collected more images (approximately 1000) with a size was larger than 1000 × 1000. The final selection included 2413 images of various sizes (between 600 × 600 and 1400 × 1400) and 34 146 objects in 7 classes. The scales of the objects widely varied from scene to scene. To test our neural network by challenging it to detect one-channel gray scale images, we merged the three-channel RGB images into one-channel gray scale images to destroy the color information. The RGB information destroyed by merging the three channels is unrecoverable; therefore, the converted-gray scale images can be regarded as one-channel gray scale images. As the corresponding dataset of the one-channel gray scale images was represented by three-channel RGB images, this allowed us to use one of the neural network's architectures to evaluate both RGB and one-channel gray scale images without constructing a new neural network.

The distribution of the dataset is shown in Table II. The dataset was distributed over seven classes: ships, aircraft, playground, viaducts, ports, storage tanks, and bridges. In our dataset, the ports, playground, viaducts, and bridges covered considerably more space in the images than the other three. Thus, the total number for these four items was less than the other three. We balanced the dataset with at least 763 objects that could be used in the training procedure to guarantee the DNN's training.

Feature size and image size are two other factors of object detection in HRSI. In our datasets, we collected 2413 images that were mostly larger than 600 × 600 to better suit HRSI detection. The descriptive statistics for the datasets are shown in Table II. The other important factor, meter/pixel, is shown in Table III. Most images containing small objects (ships, aircraft, and storage tanks) were collected from images above level 19 (<0.5 m) in Google Earth. This was conducted to maintain clarity for neural network training. As the other three items generally covered large regions on land, thus 0.5 m images composed about 40% of our dataset. To test the training results, 30% of the images were randomly selected as test and validation data. As the focus was small objects and narrow rectangle objects, we maintained their proportion at around 30% and 20% in the dataset, respectively. All objects also had different scales and different sizes. In the number of samples, the dataset did not require further dataset extension.

D. Evaluation for One-Channel Gray Scale Images

One-channel gray scale images, such as in panchromatic images, are widely used in the remote sensing community. To evaluate the performance of object detection neural networks with one-channel gray scale images, we converted the RGB dataset to gray scale with the classical three-channel RGB image formula

$$G = \frac{0.299 \times \mathrm{R} + 0.587 \times \mathrm{G} + 0.114 \times \mathrm{B}}{3} \tag{5}$$

where G denotes gray and R, G, and B means red, green, and blue respectively.

As such, a corresponding one-channel gray scale image dataset was built. Another issue was the channels for training. Because most state-of-the-art neural networks for object detection are based on three channels, training the neural network with one-channel gray scale images input does not provide a fair evaluation of their performance. Thus, we simply copied the value of the one-channel gray scale images to all three channels to represent the gray scale images with three channels for the purpose of training the neural networks. As merging the three-channel RGB images into one-channel gray scale images causes unrecoverable damage to their information, the corresponding dataset can be used to efficiently and fairly evaluate the state-of-the-art DNN detectors' performance with one-channel gray scale images.

E. Comparative Evaluation

As shown in Table IV, our neural network is the most accurate for both the three-channel RGB dataset and the one-channel gray scale dataset. These results show that our neural network more accurately localizes all seven target classes than the other methods.

According to the results shown in Fig. 10, our neural network can consistently locate the correct region and recognize the correct classification of multiple objects. Even under foggy conditions, the neural network can ideally localize the target objects with proper ROIs, as shown in the first column of Fig. 10. This indicates that our approach can overcome the challenges presented by blurring. Regarding the experiments based on VGG and ResNet as backbones, very slight difference can be found between them. This illustrates that the backbone trained by the ImageNet can only slightly influence the aerial view object detection. Furthermore, our neural network detects the small objects are hardly recognizable by the naked eye. In the selected targets such as ship, aircraft, viaduct, storage tank, and bridge are all in multi-scale. The small objects are mainly contained in

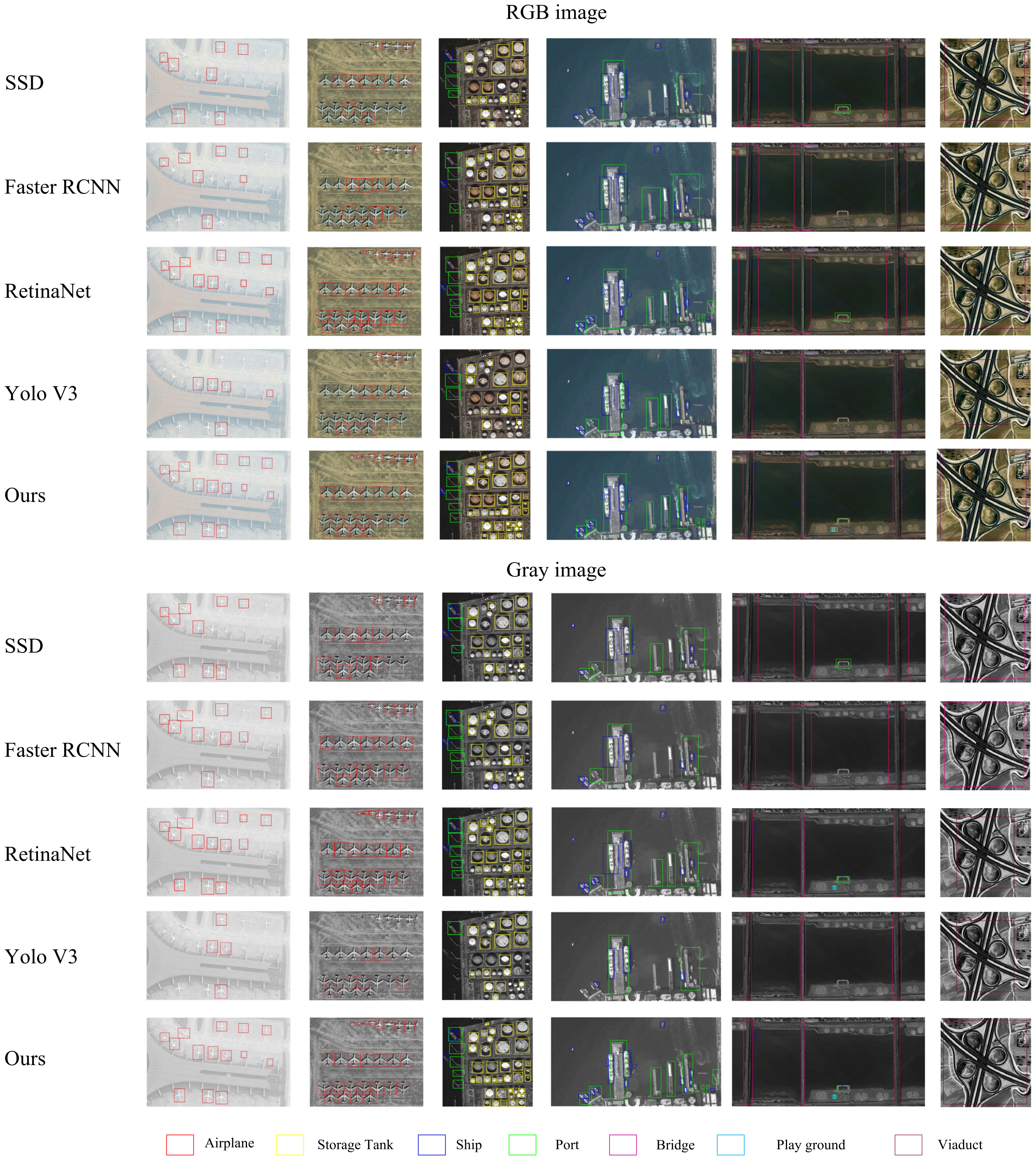

Fig. 10. Results of our proposed methods in gray and RGB imagery comparing with other state-of-the-art methods.

ship, aircraft and storage tanks. The narrow rectangular objects are mainly contained in ship and bridge.

However, the number of small-size and narrow rectangular objects are hard to be clearly defined, but they occupy around 30% and 20% in the dataset. The object detection based on our proposed method for the abovementioned targets still obtained very promising accuracy compared with other state-of-the-art methods, which proved that our proposed method can detect the objects with scale invariance ability. The major reason is that our proposed method employs the double-shot strategy, which opens an extra path for the features during extraction. The two paths play different roles in localizing and classifying the objects, and focusing on detecting different size objects through using misplaced localization strategy and normal size anchor

TABLE IV
TESTING RESULTS OF RGB AND ONE-CHANNEL DATASET

(BY RESNET)

(a) Training and Testing in Three-Channel RGB Dataset

Item	SSD	Faster RCNN	Retina Net	Yolo V3	Ours
Bridge	0.734	0.757	0.825	0.558	0.836
Aircraft	0.887	0.620	0.800	0.322	0.910
Storage Tank	0.710	0.725	0.855	0.632	0.858
Port	0.265	0.210	0.317	0.209	0.305
Ship	0.650	0.530	0.809	0.604	0.828
Play-ground	0.834	0.838	0.861	0.727	0.914
Viaduct	0.864	0.840	0.893	0.790	0.910
MaP	0.706	0.646	0.766	0.549	0.794

(b) Training and Testing in One-Channel Gray scale Dataset

Item	SSD	Faster RCNN	Retina Net	Yolo V3	Ours
Bridge	0.730	0.755	0.820	0.606	0.803
Aircraft	0.912	0.583	0.793	0.313	0.885
Storage Tank	0.736	0.659	0.882	0.582	0.840
Port	0.240	0.252	0.342	0.143	0.305
Ship	0.621	0.505	0.764	0.621	0.820
Play-ground	0.825	0.815	0.856	0.743	0.892
Viaduct	0.900	0.840	0.895	0.801	0.901
MaP	0.709	0.630	0.765	0.544	0.778

(BY VGG)

(c) Training and Testing in Three-Channel RGB Dataset

Item	SSD	Faster CNN	Ours
Bridge	0.723	0.758	0.824
Aircraft	0.898	0.620	0.902
Storage Tank	0.735	0.725	0.864
Port	0.224	0.210	0.316
Ship	0.615	0.530	0.814
Play-ground	0.814	0.838	0.903
Viaduct	0.903	0.840	0.918
MaP	0.701	0.646	0.792

(d) Training and Testing in One-Channel Gray scale Dataset

Item	SSD	Faster CNN	Ours
Bridge	0.731	0.746	0.793
Aircraft	0.902	0.576	0.902
Storage Tank	0.725	0.693	0.834
Port	0.268	0.162	0.311
Ship	0.567	0.532	0.814
Play-ground	0.830	0.806	0.890
Viaduct	0.912	0.862	0.890
MaP	0.705	0.625	0.775

TABLE V
TESTING RESULTS OF MODIFYING RGB AND ONE-CHANNEL GRAY SCALE DATASETS

(BY RESNET)

Training in RGB Dataset, Testing in One-Channel Dataset

Item	SSD	Faster RCNN	Yolo V3	Retina Net	Ours
Bridge	0.748	0.745	0.536	0.865	0.842
Aircraft	0.860	0.612	0.348	0.753	0.927
Storage Tank	0.735	0.730	0.620	0.836	0.848
Port	0.325	0.219	0.153	0.327	0.292
Ship	0.627	0.548	0.613	0.796	0.843
Play-ground	0.822	0.830	0.693	0.841	0.905
Viaduct	0.845	0.842	0.802	0.885	0.902
MaP	0.709	0.647	0.538	0.778	0.794

Training in One Chanel Dataset Testing in RGB Dataset

Item	SSD	Faster RCNN	Yolo V3	Retina Net	Ours
Bridge	0.423	0.532	0.428	0.596	0.619
Aircraft	0.547	0.425	0.283	0.692	0.767
Storage Tank	0.555	0.545	0.468	0.707	0.674
Port	0.128	0.217	0.184	0.240	0.241
Ship	0.357	0.325	0.450	0.593	0.630
Play-ground	0.635	0.578	0.563	0.691	0.761
Viaduct	0.522	0.728	0.620	0.682	0.703
MaP	0.452	0.478	0.428	0.600	0.628

(BY VGG)

Training in RGB Dataset, Testing in One-Channel Dataset

Item	SSD	Faster CNN	Ours
Bridge	0.723	0.762	0.840
Aircraft	0.898	0.591	0.937
Storage Tank	0.736	0.689	0.852
Port	0.256	0.227	0.257
Ship	0.615	0.511	0.765
Play-ground	0.802	0.818	0.895
Viaduct	0.903	0.829	1.000
MaP	0.703	0.632	0.789

Training in One Chanel Dataset Testing in RGB Dataset

Item	SSD	Faster CNN	Ours
Bridge	0.411	0.527	0.611
Aircraft	0.502	0.455	0.759
Storage Tank	0.585	0.500	0.664
Port	0.168	0.180	0.231
Ship	0.367	0.355	0.626
Play-ground	0.530	0.590	0.775
Viaduct	0.512	0.608	0.693
MaP	0.440	0.459	0.622

localization. Unlike other methods, it effectively avoids the feature crush in different scales and *h/w* size objects in the same sequence and proved its efficiency in the experiment results. This illustrates to the community that path division enables to assist the localization and classification of complicated size objects, and anchor misplaced localization helps the object detection of various *h/w* sizes. One trivial drawback is that the two paths are without connection, and potential improvement can be expected to improve the detecting accuracy in the future.

F. Impact of Gray Scale Images Mixed With RGB Images

First, we evaluated the performance of our neural network and other state-of-the-art methods. We used the networks trained by the three-channel RGB datasets to test the one-channel gray scale dataset and networks trained by the gray scale dataset to test the three-channel RGB model.

The results are shown in Table V. Each of the DNNs maintains accuracy when using the three-channel RGB model to test the one-channel gray scale datasets. However, the MaP significantly decreased 15% when using the one-channel gray scale model to test the three-channel RGB dataset. These results illustrate that the features learned through training dataset.

As shown in Table V, object detection in the one-channel gray scale image is as accurate as in the three-channel RGB datasets. This reveals that destroying the RGB information by merging three channels to one channel does not significantly influence the feature extraction strength. The results of training with different proportions between RGB images and one-channel images does not significantly influence feature extraction strength. Next, we

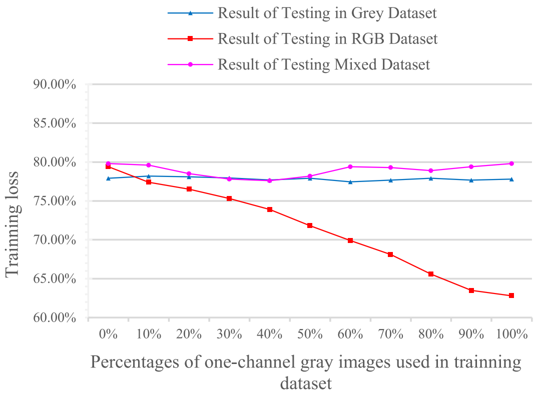

Fig. 11. Detection result of our proposed method detecting with different proportions of gray and RGB datasets.

evaluated the influence of merging the channels with the three-channel RGB images corresponding to the features of the one-channel gray scale images. However, due to the unrecoverable destruction of information during the merging of the channels, the one-channel gray scale models cannot detect objects in three-channel RGB images.

Then, we randomly mixed different percentages of the one-channel gray scale images into the three-channel RGB images. The training results, by proportion, for RGB images and one-channel images in the training dataset indicate the performance of our neural network in the context of different types of merged images. As shown in Fig. 11, the MaP for training our DNN with different percentages of one-channel gray scale images was 78.9% in tests with one-channel gray scale images. The testing accuracy in the one-channel gray scale dataset, three-channel RGB dataset, and 50% mixed dataset decreased as the percentage of one-channel gray scale images increased. First, the RGB dataset employs three channels but the gray scale image only has one channel (i.e., copying the same value to three channels.). The RGB images contain richer feature in texture information. According to the architecture of feature extraction, the low-level features are contained in the early stage of feature maps, whereas the high-level features contained in the shallow-level feature maps. As the goal of the object detection DNN is to localize the rough positions of target objects in VHR images, the proposals are mostly estimated in the shallow feature maps after the backbone. The texture information, which mainly contains the low-level features, of objects inside VHR images is less complex compared with that ground view images. The shape, edge, and features (high-level) of the objects remain after the merging process, which allows the model to be trained by gray scale imagery to detect objects in the RGB channel dataset. This proves that the high-level features play a more important role than the low-level features in the object detection on remote sensing images. In the experiments, the model trained by the gray dataset was 62.8% accurate, which is less accurate than for RGB dataset detection. With increasing RGB image proportion, the accuracy increases to 76% by including half of the RGB dataset.

These results revealed that different types of images can be mixed without influencing the detection performance, except in cases of three-channel RGB images. The poor performance when using the model trained in one-channel dataset tested on three-channel RGB images indicates that the training dataset must contain enough three-channel RGB images when used to detect targets using the DNN. Fewer and simpler features are extracted in one-channel gray scale images than in three-channel RGB images. Therefore, training and testing the neural network on one-channel gray scale images can produce reasonable accuracy. However, the method does not work with more feature-rich datasets. Therefore, the MaP for the mixed dataset decreased as the proportion of one-channel gray scale images increased.

IV. Conclusion

This article illustrates a novel DNN that detects multiple objects in Google Earth satellite images. Its ConvNet architecture is significantly more accurate than previous iterations. We compared this approach with existing methods using several experiments. Several conclusions can be drawn.

1) Our proposed neural network was the most accurate (79.4% MaP in three-channel RGB datasets and 77.8% MaP in one-channel gray scale datasets). This is at least 4% higher than other state-of-the-art DNNs.
2) In evaluation using SNRO-7, the experiments showed that our proposed neural network was the most accurate for aircraft (91.0%), storage tank (86.0%), and ships (82.4%) among other state-of-the-art methods for small object detection. The high input size and double-shot strategy allow our neural network to localize small objects, even when the naked eye can hardly recognize them.
3) Our proposed DNN obtained the highest accuracy for the narrow rectangle item bridge (83.4%) among the considered methods. In combination with the misplaced localization strategy, the neural network obtained a more precise ROI for narrow rectangle objects.
4) The evaluation for training and testing with the one-channel gray scale images illustrated that models trained by a three-channel RGB dataset can detect objects in gray scale images. However, performance significantly worsens when models trained by a one-channel gray scale for object detection on RGB images.
5) The results of training and testing the neural network by mixing RGB with gray-scale images showed that different types of images can be used together for training with limited detection performance loss. This can guide researchers during dataset selection and help minimize dataset preparation time.

References

[1] Y. Li, K. Fu, H. Sun and X. Sun, “An aircraft detection framework based on reinforcement learning and convolutional neural networks in remote sensing images,” *Remote Sens.*, vol. 10, no. 2, 2018, Art. no. 243.

[2] F. Zhang., B. Du, L. Zhang, and M. Xu, “Weakly supervised learning based on coupled convolutional neural networks for aircraft detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[3] A. Zhao *et al.*, “An effective method based on ACF for aircraft detection in remote sensing images,” *IEEE Trans. Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 744–748, May 2017.

[4] H. He, Y. Lin, F. Chen, H. Tai, and Z. Yin, “Inshore ship detection in remote sensing images via weighted pose voting,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3091–3107, Jun. 2017.

[5] H. Lin, Z. Shi, and Z. Zhou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1665–1669, Oct. 2017.

[6] J. Tang, C. Deng, G. B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1174–1185, Mar. 2015.

[7] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3652–3664, Aug. 2017.

[8] H. Schilling, D. Bulatov, R. Niessner, W. Middelmann, and U. Soergel, "Detection of vehicles in multisensor data via multibranch convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4299–4316, Nov. 2018.

[9] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X.X. Zhu, "R^3-net: A deep network for multi-oriented vehicle detection in aerial images and videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5028–5042, Jul. 2019.

[10] T. Moranduzzo and F. Melgani, "Detecting cars in UAV images with a catalog-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6356–6367, Oct. 2014.

[11] G. Cheng; and J. A. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016.

[12] X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, and F. Xu, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[13] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, 2019.

[14] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100. Jan. 2018.

[15] S. M. Azimi, P. Fischer, M. Korner, and P. Reinartz, "Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2920–2938, May 2019.

[16] L. Mou, Y. Hua, and X. X. Zhu, "Spatial relational reasoning in networks for improving semantic segmentation of aerial images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5232–5235.

[17] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[18] R. Girshick, "Fast R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[19] S. Ren., K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

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

[21] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput Vis.*, 2016, pp. 21–37.

[22] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[23] M. Everingham, S.A. Eslam, L. Van Goo, C.K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[24] M. Kang, X. Leng, Z. Lin, and K. Ji, "A modified faster R-CNN based on CFAR algorithm for SAR ship detection," in *Proc. Int. Workshop Remote Sens. Intell. Process.*, 2017, pp. 1–4.

[25] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[26] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

[27] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[28] Y. Hua, L. Mou, and X.X Zhu, "Label relation inference for multi-label aerial image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5244–5247.

[29] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, 2009, pp. 248–255.

[30] M. D. Zeiler, and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput Vis.*, 2014, pp. 818–833.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Int. Conf. Comput. Vis.*, 2016, pp. 770–778.

[32] A. Neubeck, and L. Van. Gool, "Efficient non-maximum suppression," in *Proc. Int. Conf. Pattern Recognit.*, 2006, vol. 3, pp. 850–855.

[33] T. Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput Vis.*, 2016, pp. 740–755.

[34] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, 2014.

[35] S. Rude, "An overview of gradient descent optimization algorithms," 2016, *arXiv: 106.04747v2*.

[36] S. Jiang, and Z. Hong, "Benchmark dataset for small and narrow rectangular object detection from google earth imagery," IEEE Dataport, 2019. Accessed: Jul. 30, 2019. [Online]. Available: http://dx.doi.org/10.21227/j7nx-2495

[37] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, and L. Zhang, "AID: A benchmark dataset for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

Shenlu Jiang received the B.Eng. degree in computer science and technology from the College of Information Technology, Shanghai Ocean Universty, Shanghai, China, in 2015. He is currently working toward the Ph.D. degree in the School of Electrical and Electronics Engineering, Sungkyunkwan University, Suwon, South Korea.

From October 2018 to December 2019, he was a Research Assistant with the Department of Land Surveying and Geo-Informatics, Hongkong Polytechnic University. He is currently a Research Assistant with the Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong. His research interests include computer vision, robot vison, remote sensing, and deep learning.

Wei Yao received the B.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2003, the Dipl.-Ing. (univ.) degree in geodesy and geoinformation and the Ph.D. degree from the Technische Universität München (TUM), Munich, Germany, in 2007 and 2010, respectively.

Since 2007, he has been a Scientific Collaborator and a Lecturer with the Institute of Photogrammetry and Cartography, TUM. Since 2011, he has also been as a Senior Scientist in the research cluster of computer vision, remote sensing, and navigation with Munich University of Applied Sciences, Munich, Germany. In 2017, he was selected for National Thousand Young Talents Program of China and joined the Hong Kong Polytechnic University as an Assistant Professor. His research work was also funded by Bavarian Excellence program based on the Bavarian Elite Aid Act. He has authored/coauthored nearly 90 academic articles in refereed international journals and conferences. His main research interests include active remote sensing technology toward reconstruction and analysis of spatial-temporal behaviors of objects, image processing and analysis, machine learning, and related environmental and industrial applications.

Dr. Yao was the recipient of Best Student Paper Award from 2009 IEEE/ISPRS Joint Event on Urban Remote Sensing. Meanwhile, he was named a winner of the Chinese Government Award for Outstanding Self-Financed Students Abroad, which is granted around the world across all disciplines. He was also the recipient of Best Presentation Award of the International Symposium on Mobile Mapping 2013 and best paper awards of several IEEE/ISPRS conferences. Since 2016, he has been the Co-Chair of ISPRS WG III/6.

Man Sing Wong received the M.Phil. and Ph.D. degrees in remote sensing and geographic information system from the Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, in 2005 and 2009, respectively.

During 2006–2007, he was a Fulbright Junior Scholar with the Earth System Science Interdisciplinary Center, University of Maryland, College Park. He is the Official Site Manager for the NASA's AERONET station in Hong Kong. He has been working in various projects, including the use of remote sensing to study urban heat island effect, urban environmental quality, landslides, vegetation and ecosystems, spectral mixture analysis, aerosol retrieval, and air quality and dust storm monitoring. He has authored/coauthored approximately 100 SCI journal publications since 2005.

Dr. Wong was the recipient the Early Career Award from the Hong Kong Research Grants Council in 2014, two Faculty Award for Outstanding Performance/Achievement Award in Teaching in 2014 and 2016, respectively, a Dean's Award for Outstanding Achievement in Research Funding in 2016, and a Departmental Outstanding Research Award in 2017 from the Hong Kong Polytechnic University.

Gen Li (Student Member, IEEE) received the B.S. degree in electronic information engineering from Xidian University, Xi'an, China, in 2018. He is currently working toward the M.S. degree in electronic, electrical, and computer engineering with Sungkyunkwan University, Suwon, South Korea.

His research interests include computer vision, deep learning, object detection, and semantic segmentation.

Zhonghua Hong (Member, IEEE) received the Ph.D. degree in transportation engineering (GIS) from Tongji University, Shanghai, China, in 2014.

He is currently a Lecturer with the College of Information Technology, Shanghai Ocean University, Shanghai, China. His research interests include three-dimensional damage detection, coastal mapping, photogrammetry, GNSS-R, and deep learning.

Tae-Yong Kuc received the B.S. degree in control and instrumentation engineering from Seoul National University, Seoul, South Korea, in 1988, and the M.S. and Ph.D. degrees in control and robotics from the Pohang University of Science and Technology, Pohang, South Korea, in 1990 and 1993, respectively.

From April to August 1993, he was a Chief Research Engineer with the Precision Machinery Institute, Samsung Aerospace Company. From September 1993 to February 1995, he was as a Senior Lecturer with the Department of Electrical Engineering, Mokpo National University, South Korea. Since March 1995, he has been with the School of Electrical and Electronics Engineering, Sungkyunkwan University, Suwon, South Korea, where he is currently a Professor. His research interests include intelligent robotics, adaptive and learning control, and visual sensor processing for computer-aided control systems.

Xiaohua Tong (Senior Member, IEEE) received the Ph.D. degree in transportation engineering from Tongji University, Shanghai, China, in 1999.

From 2001 to 2003, he was a Postdoctoral Researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China. He was a Research Fellow with Hong Kong Polytechnic University, Hong Kong, in 2006, and a Visiting Scholar with the University of California, Santa Barbara, CA, USA, from 2008 to 2009. His research interests include photogrammetry and remote sensing, trust in spatial data, and image processing for high-resolution satellite images.