remote sensing

MDPI

Article

$\mathcal{F}^3$-Net: Feature Fusion and Filtration Network for Object Detection in Optical Remote Sensing Images

Xinhai Ye [1], Fengchao Xiong [1,*], Jianfeng Lu [1], Jun Zhou [2] and Yuntao Qian [3]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; yyxxhh@njust.edu.cn (X.Y.); lujf@njust.edu.cn (J.L.)
[2] School of Information and Communication Technology, Griffith University, Nathan 4111, Australia; jun.zhou@griffith.edu.au
[3] Institute of Artificial Intelligence, College of Computer Science, Zhejiang University, Hangzhou 310027, China; ytqian@zju.edu.cn
* Correspondence: fcxiong@njust.edu.cn

Received: 2 November 2020; Accepted: 4 December 2020; Published: 9 December 2020

Abstract: Object detection in remote sensing (RS) images is a challenging task due to the difficulties of small size, varied appearance, and complex background. Although a lot of methods have been developed to address this problem, many of them cannot fully exploit multilevel context information or handle cluttered background in RS images either. To this end, in this paper, we propose a feature fusion and filtration network ($\mathcal{F}^3$-Net) to improve object detection in RS images, which has higher capacity of combining the context information at multiple scales while suppressing the interference from the background. Specifically, $\mathcal{F}^3$-Net leverages a feature adaptation block with a residual structure to adjust the backbone network in an end-to-end manner, better considering the characteristics of RS images. Afterward, the network learns the context information of the object at multiple scales by hierarchically fusing the feature maps from different layers. In order to suppress the interference from cluttered background, the fused feature is then projected into a low-dimensional subspace by an additional feature filtration module. As a result, more relevant and accurate context information is extracted for further detection. Extensive experiments on DOTA, NWPU VHR-10, and UCAS AOD datasets demonstrate that the proposed detector achieves very promising detection performance.

Keywords: context information; object detection; feature filtration; convolutional neural networks (CNNs); optical remote sensing image

1. Introduction

With the fast development of airborne and spaceborne sensors, remote sensing (RS) images have become widely available, offering new opportunities to observe and interpret the Earth. Object detection aims at simultaneously determining the location and category of objects of interest in the RS image. It is an important task in practical applications of RS images such as resource acquisition, disaster monitoring, urban planning, etc. [1], and has attracted a lot of interest from both academia and industry.

Generally, object detection in RS images can be accomplished by optical images [2], synthetic aperture radar (SAR) technology [3–6], and more. In our study, we focus on object detection in optical remote sensing images. In the literature, object detection in RS images can be grouped by template matching-based, knowledge-based, and machine learning-based methods [7]. Treating this task as a classification problem, machine learning-based methods stand out due to the advance of powerful feature representations and classifiers. Representative classifiers include support vector machine

(SVM) [8], Adaboost [9], and so on. In addition to classifiers, feature representation always plays very important role in machine learning-based detection. In early years, low-level handcrafted features, such as histogram of oriented gradients (HOG), and mid-level features, e.g., bag-of-words (BoW) feature, were commonly explored for detecting RS objects [10]. Recently, increasingly more attention has been focused on deep detectors with high-level learned features, driven by strong capability of deep learning models as feature extractors and easy access to large-scale RS datasets such as DOTA [2], NWPU VHR-10 [10], UCAS-AOD [11], etc.

With the proposition of several deep detectors such as Faster Region Convolutional Neural Networks (Faster R-CNN) [12], Single Shot MultiBox Detector (SSD)[13], and You Only Look Once (YOLO) [14], very impressive performance has been achieved for object detection in natural scenes, where the objects are imaged in close range. However, when applying these detectors on RS images captured by camera mounted on satellite or airplane, the detection performance drops dramatically [2,15]. As revealed in Figure 1, object detection in RS images faces the following challenges.

- **Small objects.** Due to the long distance of imaging and low spatial resolutions of sensors, RS images often contain objects with small sizes, leading to limited information of object features.
- **Appearance variance.** Objects in the same class tend to appear in arbitrary orientations, varied sizes, and sometimes, extreme aspect ratios such as bridges and boats.
- **Background complexity.** Objects are often overwhelmed by cluttered background which potentially introduces more false positives and noise.

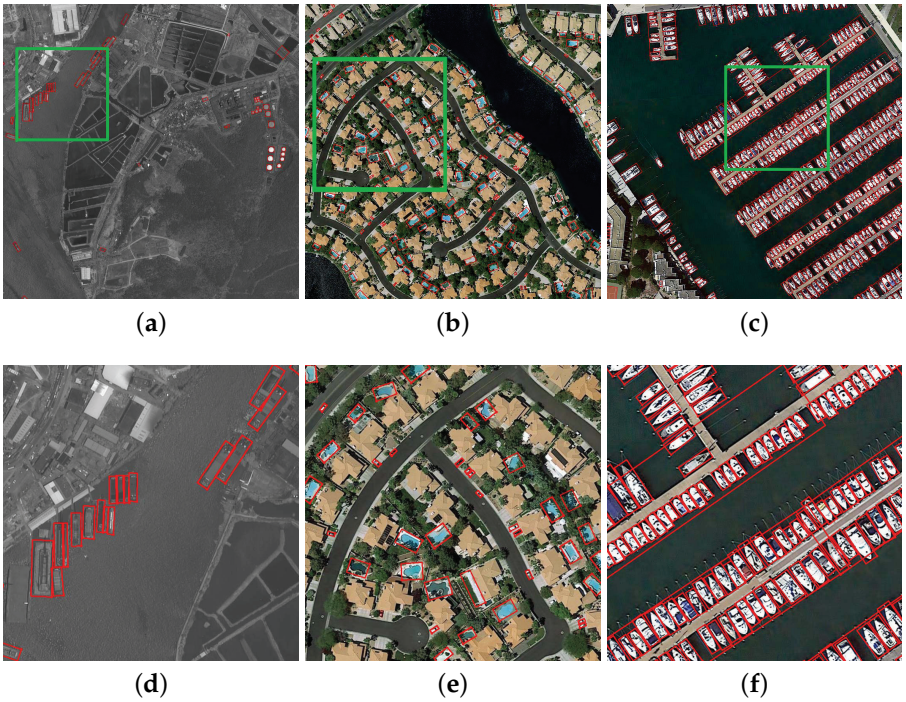


Figure 1. Illustration of detection results produced by $\mathcal{F}^3$-Net. (**a**–**f**) Thanks to the feature fusion module and feature filtration module, the objects with small sizes, dense arrangement, varied appearance, and cluttered environment are accurately detected and recognized.

To this end, many researches attempt to tackle object detection in RS images by considering the above unique characteristics of RS objects and embedding related factors into the existing networks [16–18]. For example, Cheng et al. [19] added a rotation-invariant layer to R-CNN (RICNN) to enable arbitrary oriented detection, which enforces CNN feature representations to share close mapping before and after rotation. Ding et al. [20] took the geometry transformation between horizontal and rotational RoIs into account and developed a lightweight Region of Interest (RoI)

Transformer for rotation-invariant region feature extraction. Following the rotational regional proposal network (RRPN) [21], many researchers incorporate rotation-aware factors into a regional proposal network (RPN) to handle object rotation variations [22–24]. Specifically, Li et al. [25] embedded additional multi-angle anchors into RPN for the generation of multi-scale and translation-invariant candidate regions. Xu et al. [26] and Ren et al. [27] replaced traditional convolutions with deformable convolutions to account for orientation diversity of aerial objects. Moreover, an attention mechanism is also adopted to guide the network to focus on the most irrelevant information, i.e., prominent foreground regions, which helps to mitigate the interference of cluttered background and noise [28–31].

This paper aims to address the above challenges by proposing a novel feature fusion and filtration network ($\mathcal{F}^3$-Net), whose framework is illustrated in Figure 2. As shown in the figure, $\mathcal{F}^3$-Net follows the pipeline of Faster R-CNN with the introduction of two additional modules, i.e., feature fusion module and feature filtration module, to cope with the special property of objects in RS images. The feature fusion module aims to extract the context information at different scales by combining low-resolution high-level semantic features from deeper layers and high-resolution low-level semantic features from shallow layers. This helps to address the difficulties of detecting RS objects of small sizes and appearance variance. The feature filtration block is devised to compact the fused features, which provides a novel approach to suppress irrelevant information belonging to the surrounding background and noise so that the network pays more attention to the objects of interest. This facilitates object detection in cluttered environment. Experiential results on three widely used datasets, DOTA [2], NWPU VHR-10 [10], and UCAS AOD [11] demonstrate that proposed detector is able to yield higher detection accuracy than alternatives.

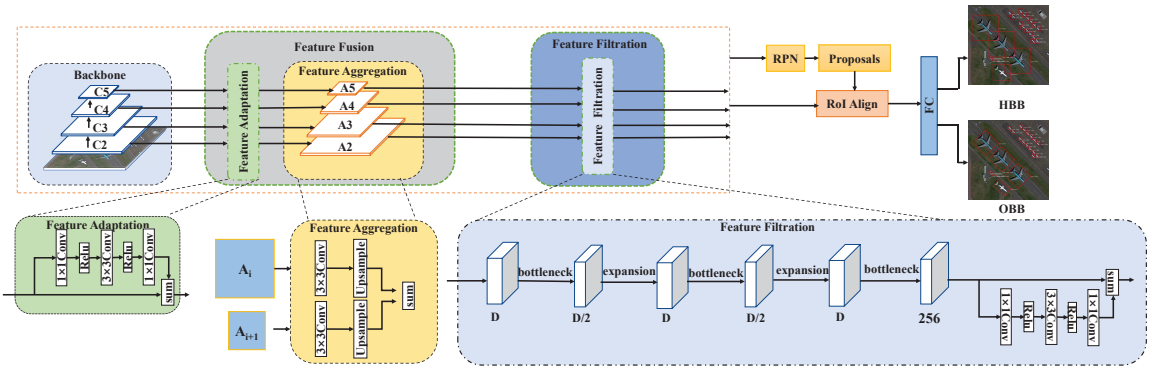


Figure 2. Framework of the proposed $\mathcal{F}^3$-Net. $\mathcal{F}^3$-Net consists of four main components: backbone for feature extraction, feature fusion module for multi-resolution multi-level feature combination, feature filtration module for compact feature representation, and detection module for object classification and regression. (A_i means the features produced by feature adaptation block.)

The rest of the paper is organized as follows. Section 2 reviews the recent advances of object detection in natural images and RS images. Section 3 describes the proposed $\mathcal{F}^3$-Net and analyzes its advantages in RS object detection. Section 4 analyzes the advantages of the proposed detector by comparing it with alternative approaches on three widely used datasets and providing ablation study. Section 5 concludes the paper with future work.

2. Related Works

In this section, we summarize the recent development of object detection and illustrate their relationships and differences in the context of natural and RS images.

2.1. Object Detection in Natural Images

Deep neural networks, especially convolutional neural networks (CNNs), have greatly boosted the object detection in natural images because of their dominant superiority of robust feature extraction

given large-scale datasets. In general, the CNN-based detectors can be divided into two categories: two-stage methods and one-stage methods. Two-stage detectors follow a pipeline of candidate region proposal, feature extraction, and classification. As the most pioneering two-stage detector, Region-based CNN (R-CNN) [32] used selective search, CNN architecture, and several support vector machines (SVMs) to fulfill the detection process. Though high performance is achieved by R-CNN, it is computationally demanding due to repetitive feature extraction required for all the proposals generated during the selective search step. Subsequently, Fast R-CNN [33] shared the feature extraction for all the proposals by introducing a region of interest (RoI) pooling layer, significantly accelerating R-CNN and increasing detection accuracy. Faster R-CNN [12] further employed regional proposal network (RPN) rather than selective search to directly generate object proposals from convolutional feature maps. Thanks to RPN, the inference time is greatly shortened, yielding cost-effective detection.

Alternatively, one-stage detectors consider the task as a regression problem. They directly estimate object candidates from a number of preset anchor boxes instead of region proposals, which further reduces the computation overhead further. Examples of one-stage detectors include RetinaNet [34], You Only Look Once (YOLO) [14], and Single Shot MultiBox Detector (SSD) [13]. The absence of region generation unavoidably causes foreground–background class imbalance during training [35]. Therefore, in the early years, one-stage detectors are inferior to two-stage detectors in accuracy. Lin et al. [34] proposed the Focal Loss to strengthen the training of fewer hard positives and suppress the overwhelming training of numerous easy negatives. This mitigates class imbalance and thus vastly improves the effectiveness of one-stage detectors. In addition to anchor-based detectors, anchor-free approaches such as CornerNet [36] and FCOS [37] are also extensively studied. Moreover, on the basis of state-of-the-art general object detectors, a number of advanced detectors are developed to allow for multi-scale detection [38,39], oriented detection [21,40,41], and more [42].

Our $\mathcal{F}^3$-Net falls into the category of two-stage detector. It is based on Faster R-CNN [12], with an additional angular offset module in the regression subnet to achieve rotational detection. Differently, the feature fusion module and feature filtration module are introduced for better consideration of appearance diversity and background complexity of the objects in RS images.

2.2. Context Information in RS Object Detection

The effectiveness of context information has been verified by many studies [43–45] in aerial object detection, especially for small objects or occluded objects. The intuition behind these works is that low-level high-resolution features from shallow layers favor localization, while high-level low-resolution features from deeper layers help classification. Therefore, the fusion of feature maps from different layers should strengthen object detection. Accordingly, feature pyramid network (FPN) [38] and deconvolutional single shot detector (DSSD) [39] take advantage of lateral connections to combine low-resolution semantically strong features with high-resolution semantically weak features, yielding enriched feature maps at all scales. Driven by the power of FPN in multi-scale detection, an atrous spatial feature pyramid (ASFP) [44] used atrous convolution layers at different rates for more effective fusion of multi-scale context information. Alternatively, image cascade network (ICN) [46] combined image cascade and FPN to allow extracting features at different levels and scales. Zhang et al. [15] proposed a context-aware detection network (CAD-Net) to integrate scene-level global semantics and object-level local contexts of objects for more consideration of low-contrast objects. Similarly, a balanced feature pyramid was introduced [47] to aggregate multi-scale multi-level features for robust localization of ships with different sizes. Liu et al. [48] enhanced YOLOv2 with oriented response dilated convolution and feature maps fusion from different layers, enabling object detection at multiple scales in complex geospatial images. SCRDet [41] employed sampling fusion network (SF-Net) to address inadequate anchor samples caused by small objects and samples anchors with a smaller stride when fusing high- and low-level feature maps with different resolutions.

Table 1 compares some typical computer vision (CV) and RS methods in terms of the way of region proposal generation, detection accuracy on DOTA dataset, and their highlights.

Table 1. Comparison of different remote sensing detectors.

	Type	Method	Region Proposal	Accuracy(%)	Highlights
CV methods	two-stage	R-CNN	selective search	-	CNN+SVM
		Fast R-CNN	selective search	-	RoI pooling layer
		Faster R-CNN	RPN	39.95	Regional proposal network
		R-FCN	RPN	30.84	Translation-invariant localization and classification
	one-stage	YOLOv2	No	25.49	The adoption of preset anchor boxes
		SSD	No	17.84	instead of region proposals
		Retina-Net	No	62.02	Focal loss overcoming foreground-background imbalance
RS methods	one-stage	R3Det	No	73.74	Feature alignment for accurate localization
		FMSSD	No	-	Atrous spatial feature pyramid
	two-stage	Faster R-CNN-O	RPN	54.13	Fine-tuned with oriented bounding box
		ICN	RPN	68.16	Image cascade and FPN
		RoI-Transformer	RPN	69.56	Rotated RoI learner for oriented objects
		CAD-Net	RPN	69.90	Global and local contexts exploitation
		SCR-Det	RPN	75.35	Sampling fusion network
		APE	RPN	75.75	Representing oriented objects with periodic vectors

Following the above-mentioned context-based approaches, our method also takes the surrounding information of objects into consideration and extracts the context information by integrating the features at all scales. The main differences between our method and these approaches are threefold:

- First, our method adapts the feature extraction backbone trained on natural images to RS images before capturing the context information. As mentioned earlier, RS images are different from nature images in many aspects. Directly fusing the feature maps yielded by the backbone trained on natural images cannot effectively characterize RS images. In contrast, our network can better explore these characteristics thanks to the additional feature adaptation step.
- Second, the number of feature channels is hierarchically reduced in the process of feature fusion instead of directly mapping to a fixed value, i.e., the feature sizes are still different with varied resolutions after fusion. In this way, our method has stronger information retention capability, enabling the network to better distinguish the target from the background.
- Third, the fused feature map is refined by a feature filter module with bottleneck structure before subsequent detection and classification. The top-down structure in the feature fusion module, which starts from the uppermost layers to earlier layers, may also introduce undesirable noise [49,50] due to the limited context information in deeper layers. This introduced unrelated information is not conducive to object detection, especially for small objects with dense arrangement. The bottleneck structure helps to suppress the influence of irrelevant information, such as clutter background and noise, and make the network focus more on the foreground regions.

3. Proposed $\mathcal{F}^3$-Net Detector

In this section, we describe in details the structure and main techniques proposed in this paper, i.e., feature fusion module and feature filtration module. Moreover, network learning is also presented.

3.1. Overall Architecture

As illustrated in Figure 2, $\mathcal{F}^3$-Net mainly consists of four components: (i) feature extraction module via the basic backbone, (ii) proposed feature fusion module for the integration of high-level semantic features and low-level visual features, (iii) proposed feature filtration module for producing more compact features and suppressing the impact of irrelevant features related to background and noise, and (iv) the class and box branch for predicting score and bounding box.

Specifically, ResNet is adopted as the backbone, whose $\{C_2, C_3, C_4, C_5\}$ layers are used for feature extraction. Next, the extracted features are fused and filtered by proposed feature fusion module and feature filtration module, generating multi-level multi-resolution feature maps. Afterward, the RPN is

applied on the feature maps to produce region proposals, which are then connected with RoI Align so that the spatial resolution of feature maps are mapped to the same for subsequent location and classification. Next, we will describe the details of proposed feature fusion module and feature filtration module.

3.2. Feature Fusion Module

As mentioned earlier, RS images suffer from insufficient information on objects due to their small sizes. The repeated subsampling operations such as convolution striding and pooling in deep CNNs gradually reduce the spatial resolution of feature maps. This means that the feature maps from deeper layers contain much fewer visual cues than those from shallow layers. Additionally, RS images usually cover larger areas of land in which the objects are of varied sizes and lack of visual cues. This requires the backbone to be equipped with strong feature extraction ability at multiple scales. Context information describes the relationship between an object and surrounding environment. It is very important to enrich the discriminative information and compensate the information loss caused by pooling and convolution operations, especially for small objects.

The feature pyramid network (FPN) exploits the context information by adopting lateral connections which merge multi-level multi-resolution features from the backbone in a top-down manner. The underlying principle behind FPN is that features from shallow layers represent low-level spatial contexts which benefit accurate location and features from deeper layers encode high-level semantics which facilitate classification. However, there are two problems when adopting FPN for RS object detection. First, FPN is directly connected to a backbone designed for classification tasks and is trained on natural image datasets. Therefore, the unique characteristics of RS images such as low spatial resolution and varied sizes have not been tackled and require feature adaptation before fusion. Second, FPN maps the number of features channels from different layers to the same number, which unavoidably leads to information loss. To this end, we design a new feature fusion module to pass low-resolution semantic features to high-resolution context information for more accurate detection.

Inspired by RefineNet [51], our feature fusion module includes two parts: a feature adaptation block and a feature aggregation block. The feature adaptation blocks aim to bridge the gap between the natural image datasets and RS datasets in an end-to-end manner so that the network trained on natural image datasets can better handle the high complexity of RS images. Our feature adaptation block shares a residual structure that is mathematically defined as

$$\mathbf{y}_l = \mathcal{F}(\mathbf{x}_l, \{\mathbf{W}_i\}) + \mathbf{x}_l, \tag{1}$$

where $\mathbf{x}_l$ is the input feature vector of the layers considered and $\mathbf{y}_l$ is the resulted feature vector. $\mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\})$ denotes the feature adaptation to be learned and has three layers, i.e.,

$$\mathcal{F} = \mathbf{W}_3\sigma(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}_l)). \tag{2}$$

Here, σ represents ReLU function, $\mathbf{W}_1$ and $\mathbf{W}_3$ are 1×1 convolutions for less training and inference time with less parameters, and $\mathbf{W}_2$ is a convolution operator of 3×3 in size, whose filter numbers are the same as the input. We put two feature adaptation blocks behind $\{C_2, C_3, C_4, C_5\}$ layers of the backbone in our implementation. Thanks to the skip connection in the feature fusion block, the unique characteristics of RS images can be easily propagated in the network, which facilitates the network capturing the complex relationships between land cover concepts.

After feature adaptation, we use feature aggregation block to simultaneously enrich the semantic features from shallower layers and augment the spatial contexts in deeper layers. Similar to FPN, we fuse the feature maps in a top-down manner. More specifically, feature aggregation combines feature representations from different layers by

$$\phi_l = \mathrm{Agg}(\mathbf{y}_{l-1}, \mathbf{y}_l), \tag{3}$$

where Agg(.) is the fusion function between the $(l-1)$-th and the l-th layers. We implement this function by

$$A_l = \text{Upsample}\left(\text{Conv}(\mathbf{y}_{l-1})\right) + \text{Upsample}\left(\text{Conv}(\mathbf{y}_l)\right). \quad (4)$$

Here, Conv is a set of 3×3 convolutions applied on the input features, resulting in the same number of feature channels, which is set to the smaller dimension from the inputs. The Upsample function maps the spatial resolution of convolved features to the larger resolution of the inputs. After feature fusion block, we can get the aggregated feature maps, i.e., $\{A_2, A_3, A_4, A_5\}$ in Figure 3. As can be seen, the feature maps generated by the proposed feature fusion module contain more discriminative contexts than those produced by FPN.

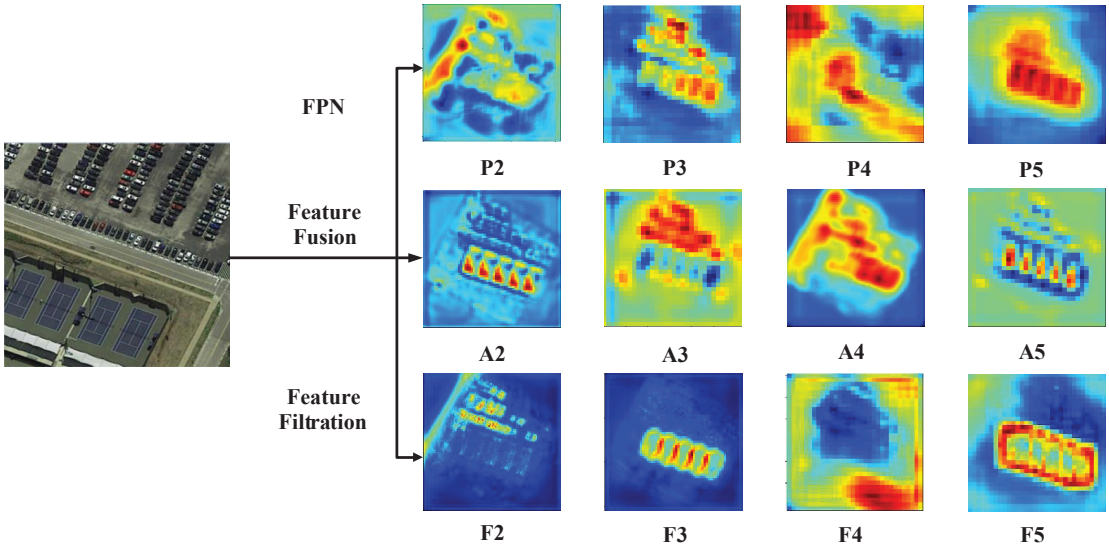


Figure 3. Visualization of feature maps produced by FPN ($\{P_2, P_3, P_4, P_5\}$), feature fusion module ($\{A_2, A_3, A_4, A_5\}$), and feature filtration module ($\{F_2, F_3, F_4, F_5\}$). For an image containing two categories, i.e., tennis court and small vehicle, feature fusion module can yield more precise context information than FPN. The feature filtration module is able to make the network focus more on the targeted objects, facilitating more accurate detection.

3.3. Feature Filtration Module

RS objects occupy only a small area of the captured image and are easily overwhelmed by the cluttered background, which potentially introduces more false positives and noise. Feature fusion module summarizes the features from different layers instead of selecting features for fusion, possibly introducing too much irrelevant information [49]. This redundant information might provide harmful cues to the detector by producing uninformative gradients and mislead object detection in RS images, especially when it comes to small objects containing only 10–20 pixels such as small vehicles. Moreover, the top-down layer-by-layer summation may also bring additional noise in the fused feature maps [50] due to limited context information in top layers. Therefore, it is improper to directly use the fused features for RS object detection. Instead, the produced features should be refined prior to subsequent classification and localization.

As illustrated in [52], given a feature map tensor $\mathcal{X}_l$ from the l-th layer containing $H_l \times W_l$ elements with D_l dimensions, the encoded information actually lies in a low-dimensional subspace. In machine learning, dimensionality reduction is a common practice to project the high-dimensional data into a low-dimensional subspace for compact feature representation, thus suppressing learning from irrelevant attributes. In neural networks, bottleneck layers with fewer neural units than input dimensions can create a compact network so that the most informational features are represented in the low dimension. Bottleneck layers can be most commonly found in autoencoder seeking a maximally compressed representation of the input feature while recovering the input information as

much as possible. Inspired by this idea, we capture the low-dimensional nature of feature maps by embedding the bottleneck structure into the convolution blocks so as to filter the negative effect of cluttered background and noise.

As shown in Figure 2, the bottleneck block includes a bottleneck layer and an expansion layer, i.e.,

$$\mathcal{F}(\mathcal{X}) = [\mathcal{G} \circ \mathcal{H}]\mathcal{X}_l. \tag{5}$$

Here, $\mathcal{G}$ is a channel-wise linear transformation to reduce dimension $\mathbb{R}^{H_i \times W_l \times D_l} \rightarrow \mathbb{R}^{H_i \times W_l \times K_l}$, and $\mathcal{H}$ is also a linear transformation to expand the low-dimensional input to a higher-dimensional space: $\mathbb{R}^{H_l \times W_l \times K_l} \rightarrow \mathbb{R}^{H_l \times W_l \times D_l}$. In our network, $\mathcal{G}$ is implemented by $\frac{D_l}{2}$ convolutions of size 1×1 and $\mathcal{H}$ is implemented by D_l convolutions of size 3×3 with the padding of 2. Considering that the side effect of nonlinearity may destroy information in low-dimensional space [52], the nonlinear activations are not inserted between two operators. With the use of two sequential bottleneck blocks, object-related features are captured while the interference of irrelevant features is suppressed. Next, the filtered feature maps are further projected to 256 channels for subsequent processing. Similar to the work in [51], a residual block is added to feature filtration block to perform non-linearity operations on the filtered feature. The dimension of the feature maps remains unchanged after this block. The feature maps generated by feature filtration module is visualized in Figure 3, i.e., $\{F_2, F_3, F_4, F_5\}$. It can be observed that irrelevant features are greatly reduced and the most informative parts are kept.

3.4. Loss Function

The same as Faster R-CNN [12], our proposed detector uses a multi-task loss defined as

$$L = \frac{\lambda_1}{N_{reg}} \sum_n p_n^* L_{reg}(t_n, t_n^*) + \frac{\lambda_2}{N_{cls}} \sum_n L_{cls}(p_n, p_n^*), \tag{6}$$

where N_{reg} and N_{cls}, respectively, denote the mini-batch size and the number of anchors. n indexes the bounding box and p_n represents the predicted probability of anchor n being an object. t_n^* and t_n are, respectively, the parameterized coordinate vectors of ground-truth and predicted bounding box. p_n^* is a binary value indicating if the anchor belongs to background ($p_n^* = 1$ for foreground and $p_n^* = 0$ for background, no regression for background). The classification loss L_{cls} is a softmax cross entropy loss across all categories. The location regression loss L_{reg} is a smooth L_1 loss defined as

$$\text{smooth}_{L_1} x = \begin{cases} 0.5x^2, & \text{if } |x| \leq 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}. \tag{7}$$

The regression of the bounding box is given by

$$\begin{aligned} &t_x = (x - x_a)/w_a,\ t_y = (y - y_a)/h_a, \\ &t_w = \log(w/w_a),\ t_h = \log(h/h_a), \\ &t_\theta = \theta - \theta_a, \\ &t_x^{'} = (x^{'} - x_a)/w_a,\ t_y^{'} = (y^{'} - y_a)/h_a, \\ &t_w^{'} = \log(w^{'}/w_a),\ t_h^{'} = \log(h^{'}/h_a), \\ &t_\theta^{'} = \theta^{'} - \theta_a, \end{aligned} \tag{8}$$

where x, y, w, h, and θ denote the center, width, height, and angle of the box, respectively. Variables x, x_a, and x^* are for the predicted box, anchor box, and ground-truth box, respectively (likewise for y, w, h).

4. Experiments

In this section, we compare the proposed $\mathcal{F}^3$-Net detector with the state-of-the-art methods, including both one-stage methods and two-stage methods, to demonstrate its advantages. The details on training and ablation study are also addressed.

4.1. Datasets and Evaluation Metric

Three widely used public datasets were adopted for experimental evaluation, including DOTA [2], UCAS AOD [11], and NWPU VHR-10 [10].

DOTA: DOTA [2] is one of the largest object detection datasets recently published for aerial images. The dataset is labeled with oriented bounding boxes (OBBs) for two tasks, i.e., horizontal bounding box (HBB) task and OBB task. It contains 2806 aerial images with size ranging from 800 × 800 to 4000 × 4000 pixels. The fully annotated dataset includes 15 categories of objects covering 188,282 instances in total. These 15 categories include plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). It is a challenging dataset due to difficulties of arbitrary scales, varied orientations, and shapes. The DOTA dataset is divided into three subsets for training (1/2), validation (1/6), and testing (1/3), respectively.

NWPU VHR-10: NWPU VHR-10 [10] is a 10-class RS object detection dataset labeled with HBB, including plane, ship, storage tank, basketball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. This dataset contains 800 aerial images, where 650 of them are labeled. Following a commonly used protocol [10], we divided the labeled images into 20% for training, 20% for validation, and the rest for testing.

UCAS AOD: UCAS AOD [11] consists of 1510 aerial images, each of which approximately covers 1000 × 1000 pixels. The dataset contains 14,596 instances of planes and cars. Same as the DOTA dataset, it covers OBB and HBB tasks. Following Xia et al. [2], 1110 images containing 700 aircraft and 410 cars were randomly selected for training, and 400 images containing 300 aircraft and 100 cars were used for testing.

The mean average precision (mAP) is adopted to evaluate the detection performance of all the methods. Formally, mAP is defined as

$$\text{mAP} = \frac{1}{C}\sum_{i=1}^{C}\int P_i(R_i)dR_i, \tag{9}$$

where R_i represents the recall for a given class i of a detector, $P_i(R_i)$ denotes the precision for a given class i when the recall of this class is R_i and C is the number of classes to be detected.

4.2. Implementation Details

Here, we describe the implementation issues of the proposed detector, including dataset preprocessing and network setup.

4.2.1. Dataset Preprocessing

Before training and testing, we first divided the original images into multiple smaller images of 800 × 800 pixels with an overlap of 600 pixels using the development kit provided by Xia et al. [2]. In this way, the GPU overhead is greatly reduced. Moreover, dataset augmentation was also adopted in the training step. We resize the original images into five scales (1024, 1024), (900, 900), (800, 800), (700, 700), and (600, 600) to apply multi-scale training in training stage. Furthermore, we augmented the data samples by rotated and flipping the images horizontally and vertically. Each image is randomly flipped with a probability of 0.5 and randomly rotated an angle from an angle set of 0°, 90°, 180°, 270°in training step. As a result, the diversity and richness of the dataset is enriched, reducing

the detection difficulty for rare categories such as bridges, helicopters, etc. The divided images were also resized into multiple sizes to allow for multi-scale detection ability.

4.2.2. Network Setup

We used ResNet [53] as our backbone. The same as Faster R-CNN [12], λ_1, and λ_2 were set to 1. The detector network was trained on a Linux machine equipped with three NVIDIA Titan XP GPUs and 12GB memory. The network was optimized using stochastic gradient (SGD) with momentum, whose weight decay was set to 0.00001. The batch size and effective mini-batch were set to 1 and 3, respectively. For DOTA and UCAS AOD datasets, the learning rate was set to 0.001 and decayed by a factor of 10 when the number of iterations reaches 720,000 and 960,000. In terms of the NWPU VHR-10 dataset, the learning rate was set to 0.0001.

4.3. Experimental Results

4.3.1. DOTA Dataset

We compared our proposed $\mathcal{F}^3$-Net with 10 state-of-the-art approaches, including one-stage methods, such as SSD [13], YOLOV2 [54], and FMSSD [44], and two-stage methods, for example, SCRDet [41], FR-O [2], ICN [46], RoI-Transformer [20], and APE [24]. Tables 2 and 3 report the quantitative comparison between the proposed $\mathcal{F}^3$-Net and several state-of-the-art approaches on both OBB and HBB tasks, respectively. For the OBB task, our $\mathcal{F}^3$-Net wins all the competing methods by achieving an mAP of 76.02%. Compared with the second best detector, i.e., APE [24], our method achieves improvement in many categories: 54.62% versus 53.42% for BR, 77.52% versus 77.16% for LV, 87.54% versus 79.45% for SH, 87.64% versus 87.15% for BC, 85.63% versus 84.51%, 64.53% versus 60.33% for RA, 78.06% versus 74.61% for HA. and 72.36% versus 71.84% for SP. In general, some categories such as LV, SV, and SH can benefit more from our feature fusion block, as they are usually very small and require more context information for accurate localization and recognition. HA and SP categories often appear in complex scenes, which interferes the detector by introducing a large number of false positives. Thanks to the feature refinement enabled by feature filtration module, the negative impact of the surrounding environment is greatly suppressed, making our detector surpass APE. In terms of HBB task, our detector also ranks the first by providing 76.48% mAP. The main reason is that our method can get more useful information than other detectors while suppressing the side effect of surrounding environment. Overall, our detector obtains very competitive detection accuracy with respect to both OBB task and HBB task.

Table 2. Experimental comparison of the baselines and our $\mathcal{F}^3-$Net for OBB task on DOTA test set.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP(%)
Two-stage methods																
FR-O [2]	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.4	52.52	46.69	44.80	46.30	52.93
ICN [46]	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.20	53.64	62.90	67.02	64.17	50.23	68.16
RoI-Transformer [20]	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
SCRDet [41]	**89.98**	80.65	52.09	68.36	68.36	60.32	72.41	**90.85**	**87.94**	**86.86**	**65.02**	**66.68**	66.25	68.24	65.21	72.61
APE [24]	89.96	**83.62**	53.42	**76.03**	**74.01**	77.16	79.45	90.83	87.15	84.51	67.72	60.33	74.61	71.84	65.55	75.75
One-stage methods																
SSD [13]	39.83	9.09	0.64	13.18	0.26	0.39	1.11	16.24	27.57	9.23	27.16	9.09	3.03	1.05	1.01	10.59
YOLOV2 [54]	39.57	20.29	36.58	23.42	8.85	2.09	4.82	44.34	38.25	34.65	16.02	37.62	47.23	25.19	7.45	21.39
R^3Det [55]	89.49	81.17	50.53	66.10	70.92	**78.66**	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	**67.17**	73.74
$\mathcal{F}^3$-Net	88.89	78.48	**54.62**	74.43	72.80	77.52	**87.54**	90.78	87.64	85.63	63.80	64.53	**78.06**	**72.36**	63.19	**76.02**

Table 3. Experimental comparison of the baselines and our $\mathcal{F}^3$−Net for HBB task on DOTA test set.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP(%)
Two-stage methods																
FR-H [2]	80.32	77.55	32.86	68.13	53.66	52.49	50.04	90.41	75.05	59.59	57.00	49.81	61.69	56.46	41.85	60.46
ICN [46]	89.97	77.71	53.38	73.26	**73.46**	65.02	78.22	90.79	79.05	84.81	57.20	62.11	73.45	70.22	58.08	72.45
SCRDet [41]	**90.18**	**81.88**	55.30	73.29	72.09	**77.65**	78.06	**90.91**	82.44	**86.39**	**64.53**	63.45	75.77	78.21	60.11	75.35
One-stage methods																
SSD [13]	57.85	32.79	16.14	18.67	0.05	36.93	24.74	81.16	25.10	47.47	11.22	31.53	14.12	9.09	0.00	29.86
YOLOV2 [54]	76.90	33.87	22.73	34.88	38.73	32.02	52.37	61.65	48.54	33.91	29.27	36.83	36.44	38.26	11.61	39.20
FMSSD [44]	89.11	81.51	48.22	67.94	69.23	73.56	76.87	90.71	82.67	73.33	52.65	**67.52**	72.37	**80.57**	60.15	72.43
$\mathcal{F}^3$-Net	88.91	78.50	**56.20**	**74.43**	73.00	77.53	**87.72**	90.78	**87.64**	85.71	64.27	63.93	**78.70**	74.00	**65.85**	**76.48**

Figure 4 visualizes the detection results for both OBB and HBB tasks on DOTA dataset. Even in the extremely complex scene, our detector can precisely get the location and accurately identify the categories (see Figure 4a,e) thanks to the merits of feature filtration module in irrelevant information suppression. The categories with small sizes and arbitrary rotation can get benefit from the feature fusion module in multi-scale multi-level feature combination, leading to the favorable results in Figure 4b,f. The remaining figures show that our detector plays well on objects with dense arrangement and low resolution with low-contrast visual cues. The promising results further prove the effectiveness of our method in RS object detection.

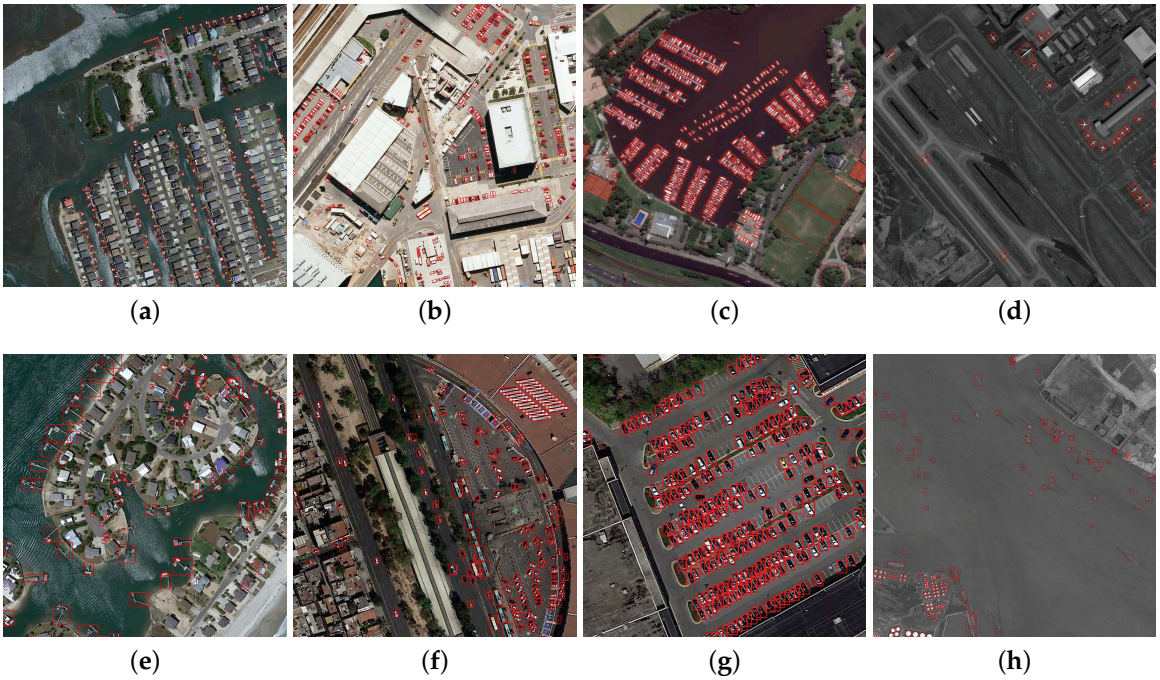


Figure 4. Examples of our detection results for both OBB and HBB tasks on DOTA test set. (**a**–**d**): OBB task and (**e**–**h**): HBB task.

4.3.2. NWPU VHR-10 and UCAS AOD Datasets

Table 4 gives comparative mAPs of all the competing methods on NWPU VHR-10 dataset. The proposed $\mathcal{F}^3$-Net provides 91.89% mAP and is the best out of all other alternative detectors. Noticeable achievement of 92.62% is obtained for the categories of ship and 91.38% for TC, which is inline with the results on DOTA dataset. As we know, ships have extreme aspect ratios, which are

easily interfered with by the irrelevant information from cluttered background. Thanks to the feature filtration module, this information is suppressed, helping the detector focus more on the predominate foreground regions. We also compared our method against ICN [46] and YOLOv2 on UCAS AOD dataset, whose results are shown in Table 5. As expected, $\mathcal{F}^3$-Net also outperformed the alternative methods by respectively giving 96.03% and 96.90% mAP scores on OBB and HBB tasks. Figure 5 gives the qualitative detection results of our method on NWPU VHR-10 and UCAS AOD datasets. Though small, densely arranged, and with cluttered background, the objects can be well located. We further plot the the precision-recall curves and F-measure curves on NWPU VHR-10 dataset in Figures 6 and 7. In summary, the excellent performance on the two datasets once again evidently demonstrate the superiority of our method.

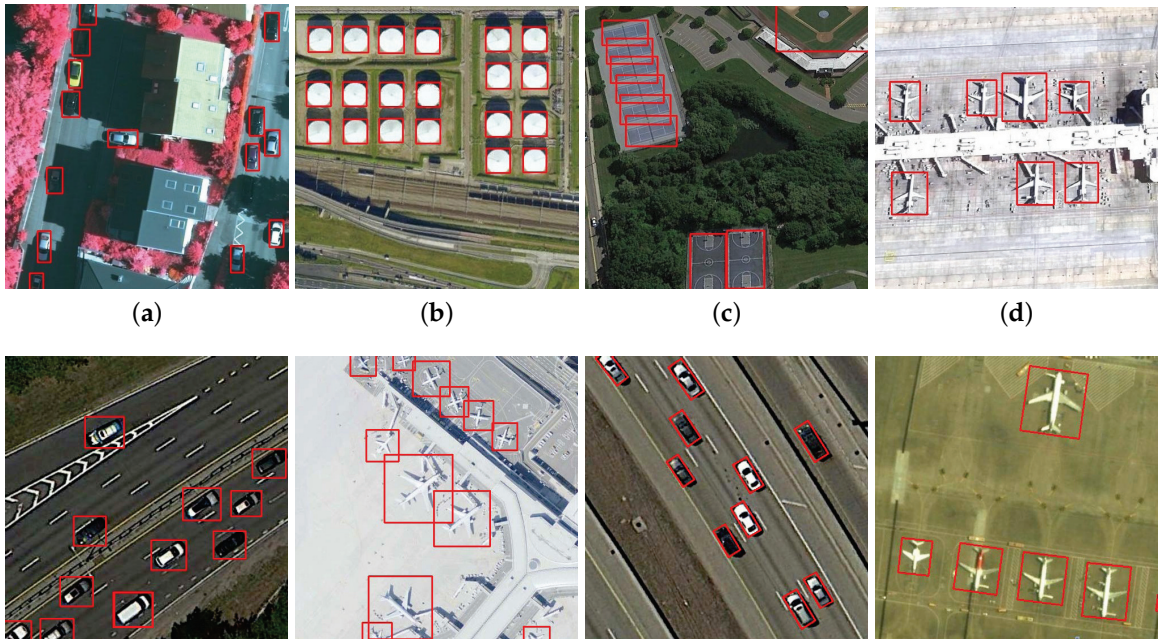


Figure 5. The visual results on NWPU VHR-10 and UCAS AOD datasets. (**a**–**d**) HBB task for NWPU VHR-10; (**e**–**f**) HBB task for UCAS AOD; (**g**–**h**) OBB task for UCAS AOD.

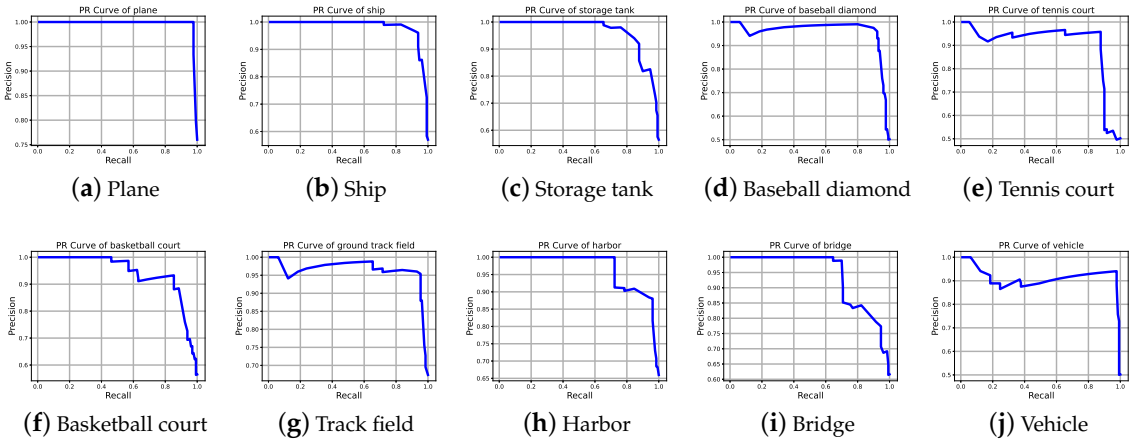
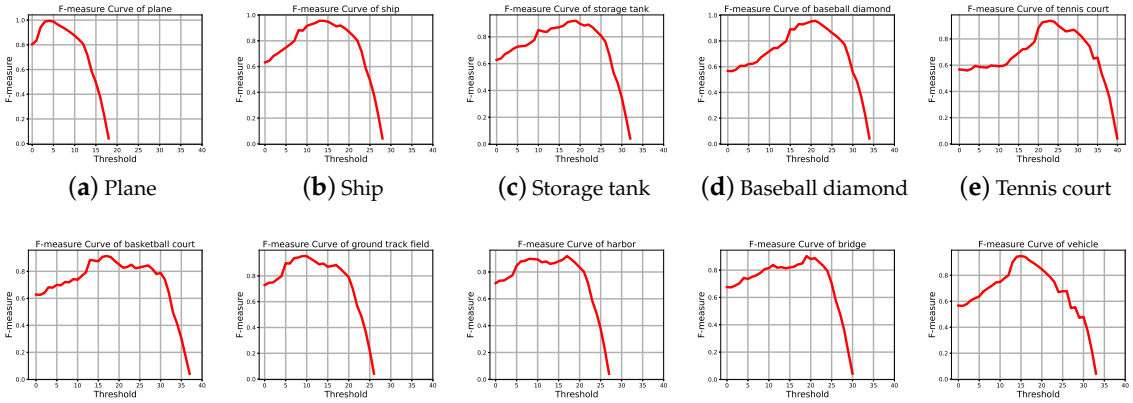


Figure 6. PR curves of different objects on NWPU VHR-10 dataset with threshold set to 0.5.



(**a**) Plane (**b**) Ship (**c**) Storage tank (**d**) Baseball diamond (**e**) Tennis court

(**f**) Basketball court (**g**) Track field (**h**) Harbor (**i**) Bridge (**j**) Vehicle

Figure 7. F-measure curves of different objects on NWPU VHR-10 dataset.

Table 4. Result comparison of baselines and $\mathcal{F}^3$-Net for HBB task on NWPU VHR-10 dataset.

Method	Plane	Ship	ST	BD	TC	BC	GTF	Harbor	Bridge	Vehicle	mAP(%)
RICNN [19]	88.35	77.34	85.27	88.12	40.83	58.45	87.63	68.60	61.51	71.10	72.63
Deformable R-FCN [26]	87.30	81.40	63.60	90.40	81.60	74.10	90.30	75.30	71.40	75.50	79.10
Deformable Faster R-CNN [27]	90.70	87.10	70.50	89.50	89.30	87.30	97.20	73.50	69.90	88.80	84.40
Li et al. [25]	**99.70**	90.80	90.60	92.90	90.30	80.10	90.80	80.30	68.50	87.10	87.10
FMSSD [44]	**99.70**	89.90	90.30	**98.20**	86.00	**96.08**	**99.60**	75.60	80.10	88.20	90.40
$\mathcal{F}^3$-Net	99.31	**92.62**	**92.89**	97.14	**91.38**	86.16	98.00	**90.30**	**82.18**	**88.90**	**91.89**

Table 5. Result comparisons on UCAS AOD dataset.

Task	Method	mAP (%)	Plane	Car
OBB	ICN [46]	95.67	-	-
	Ours	**96.03**	**98.14**	**93.92**
HBB	Xia et al. [2]	89.41	90.66	88.17
	Ours	**96.90**	**98.12**	**95.68**

4.4. Ablation Study

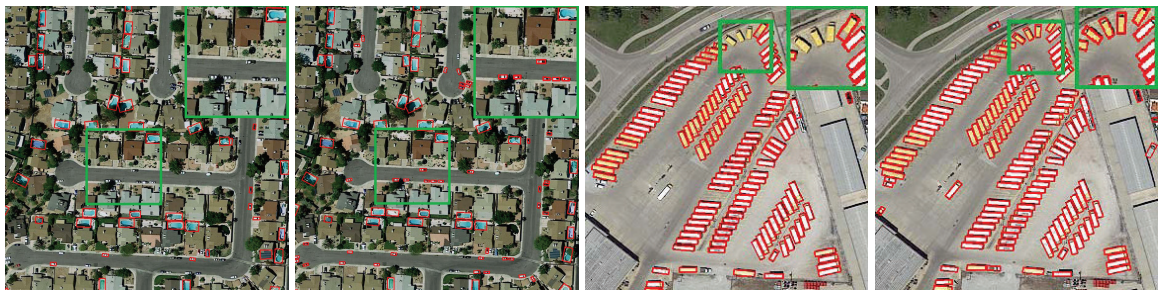
In this section, we conduct an ablation study to illustrate the effect of different modules or settings as well as their combination over the DOTA dataset. As shown in Table 6, we consider five factors that may influence the detection accuracy: feature fusion module, feature filtration module, data augmentation, backbone network setting, and multi-scale setting.

Table 6. Ablation study of components on DOTA dataset.

Faster RCNN	Backbone Network	FPN	Feature Fusion	Feature Filtration	Data Augmentation	Multi-Scale	mAP (%) @OBB	mAP (%) @HBB
✓	ResNet-50	✓	-	-	-	-	69.35	71.32
✓	ResNet-50	-	✓	-	-	-	70.87 (↑**1.52**)	72.03 (↑**0.71**)
✓	ResNet-50	-	-	✓	-	-	70.96 (↑**1.61**)	72.19 (↑**0.87**)
✓	ResNet-50	-	✓	✓	-	-	72.23 (↑**2.88**)	73.02 (↑**1.70**)
✓	ResNet-101	-	✓	✓	✓	-	73.14 (↑**3.79**)	74.62 (↑**3.30**)
✓	ResNet-152v1d	-	✓	✓	✓	-	74.26 (↑**4.91**)	75.03 (↑**3.71**)
✓	ResNet-152v1d	-	✓	✓	✓	✓	76.02 (↑**6.67**)	76.48 (↑**5.16**)

Baseline: The baseline network is a Faster-RCNN with FPN using ResNet-50 backbone. Except for the extended regression for OBB task, no modifications have been made. In our implementation, the baseline mAPs for the OBB task and HBB task are 69.35% and 71.32%, respectively.

Effect of Feature Fusion: The feature fusion module aims to enrich the feature representation for context exploitation and adapt the backbone network for RS images. The experimental results reported in Table 6 show that feature fusion module leads to a gain of 1.52% on OBB task and 0.71% on HBB task, when compared with the backbone. Figure 8 presents the detection results with this module and FPN. It can observed from Figure 8a that some very small cars are missed. Thanks to the feature fusion module, most of them are detected in Figure 8b. The visual comparison between Figure 8c and Figure 8d shows that objects with arbitrary rotation also benefit from this module. All these phenomena demonstrate that the proposed feature fusion module has more ability of exploiting the context information contained in the scene.



(**a**) (**b**) (**c**) (**d**)

Figure 8. The effectiveness of feature fusion module. Panels (**a**,**c**) show the detecting results with FPN and panels (**b**,**d**) provide the results with feature fusion module.

Effect of Feature Filtration: The feature filtration module is proposed to reduce the effect of cluttered background and noise by using the bottleneck structure. Table 6 shows that this module can improve the mAP by 1.61% on OBB task and 0.87% on HBB task than the baseline, respectively. This evidently shows the effectiveness of the feature filtration block in aerial object detection. As shown in Figure 9, the detector without feature filtration is prone to falsely detect the objects in cluttered environment, e.g., the detector mistakes the container as car in Figure 9a and fail to detect all the harbors in Figure 9c. The main reason is that complex environment is highly likely to introduce false positives that disturb information, for example, the reflection of water surface in Figure 9c. By using introduced feature filtration module, this negative information is greatly suppressed, resulting more reliable detection in Figure 9b,d. Moreover, the combination of feature fusion and feature filtration provides an mAP of 72.23% on OBB task and 73.02% on HBB task, which further shows their effectiveness for RS object detection.

(**a**) (**b**) (**c**) (**d**)

Figure 9. The effectiveness of feature filtration module. Panels (**a**,**c**) show the detecting results without feature filtration module and panels (**b**,**d**) provide the results with this module.

Effect of Data Augmentation The purpose of data augmentation is to increase the number and diversity of training samples. Benefiting from image rotation, flipping, and more, the mAP is increased to 73.14% on OBB task and 74.62% on HBB task, respectively. There is a 3.79% and 3.30% improvement over the baseline, respectively.

Effect of Backbone: The backbone network also plays an important role in object detection. A deeper backbone generally indicates more capacity of feature extraction. To this end, we replace the backbone network with ResNet152v1d (Link: https://gluon-cv.mxnet.io/model_zoo/classification.html). Table 6 shows that a deeper network does improve detection performance by producing detection accuracy of 74.26% and 75.03% on OBB and HBB tasks.

Effect of Multi-Scale Setting: The multi-scale training and testing is a useful tool to address the size variation issue in RS object detection. Thanks to multi-scale setting, the mAP, respectively, reaches 76.02% and 76.48% on OBB and HBB tasks.

In general, the experimental results in the ablation study show that all the introduced modules and settings are well aligned with our initial motivation.

Limitations: As can be seen in Tables 2 and 3, the limitations of our method lies in two aspects. First, our method is not good at detecting objects with similar visual appearance but different categories, for example, ground track field (GTF) and soccer ball field (SBF), which have always been the difficulties in remote sensing object detection. Second, our detector fails to accurately locate the objects with extreme aspect ratio, such as bridge (BD) and harbor (HA). The main reason is that the range of anchor size is set the same for all the objects, partly ignoring the prior shape information of the objects to be detected.

5. Conclusions

In this paper, a $\mathcal{F}^3$-Net is introduced for object detection in RS images, which captures the context information by feature fusion and feature filtration modules. The feature fusion module includes feature adaptation block and feature aggregation block, which, respectively, adapts the backbone network for more effective feature extraction and fuses the multi-level multi-scale features for context exploitation in an end-to-end manner. The bottleneck structure in feature filtration module encourages learning from irrelevant features and suppressing learning from cluttered background, which improves the detecting ability in complex scenarios. Extensive experiments on three widely used datasets show the effectiveness of $\mathcal{F}^3$-Net in RS object detection. In order to overcome the shortcomings of our detector, we will embed the fine-grained image recognition module to the detector to enhance classification objects with similar appearance. Moreover, we will integrate the prior knowledge of the objects to the network so that adaptive anchor sizes are produced, especially for objects with extreme sizes.

Author Contributions: All authors contributed to this manuscript: Conceptualization, X.Y. and F.X.; Methodology, X.Y. and F.X.; Supervision, F.X. and J.L. ; Validation, X.F and J.Z.; Resources, J.L.; Founding acquisition, J.L., F.X., and Y.Q.; Writing—original draft, X.Y.; Writing—review and editing, F.X., J.L., J.Z., and Y.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported in part by the National Major Program for Technological Innovation 2030—New Generation Artificial Intelligence of China under Grant 2018AAA0100500, the National Key Research and Development Program of China under Grant 2017YFB1300205 and 2018YFB0505000, Jiangsu Provincial Natural Science Foundation of China under Grant BK20200466, and the National Natural Science Foundation of China under Grant 62002169 and 62071421.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dong, Z.; Wang, M.; Wang, Y.; Zhu, Y.; Zhang, Z. Object Detection in High Resolution Remote Sensing Imagery Based on Convolutional Neural Networks With Suitable Object Scale Features. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2104–2114. [CrossRef]
2. Xia, G.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
3. Braga, A.M.; Marques, R.C.; Rodrigues, F.A.; Medeiros, F.N. A median regularized level set for hierarchical segmentation of SAR images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1171–1175. [CrossRef]
4. Ciecholewski, M. River channel segmentation in polarimetric SAR images: Watershed transform combined with average contrast maximisation. *Expert Syst. Appl.* **2017**, *82*, 196–215. [CrossRef]
5. Jin, R.; Yin, J.; Zhou, W.; Yang, J. Level set segmentation algorithm for high-resolution polarimetric SAR images based on a heterogeneous clutter model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 4565–4579. [CrossRef]
6. Lang, F.; Yang, J.; Yan, S.; Qin, F. Superpixel segmentation of polarimetric synthetic aperture radar (sar) images based on generalized mean shift. *Remote Sens.* **2018**, *10*, 1592. [CrossRef]
7. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
8. Moustakidis, S.; Mallinis, G.; Koutsias, N.; Theocharis, J.B.; Petridis, V. SVM-Based Fuzzy Decision Trees for Classification of High Spatial Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 149–169. [CrossRef]
9. Aytekin, O.; Zöngür, U.; Halici, U. Texture-Based Airport Runway Detection. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 471–475. [CrossRef]
10. 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. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
11. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the 4th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
15. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
16. Cheng, G.; Han, J.; Zhou, P.; Xu, D. Learning Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection. *IEEE Trans. Image Process.* **2019**, *28*, 265–278. [CrossRef] [PubMed]
17. Zhou, C.; Zhang, J.; Liu, J.; Zhang, C.; Shi, G.; Hu, J. Bayesian Transfer Learning for Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *58*, 7705–7719. [CrossRef]
18. Dong, R.; Xu, D.; Zhao, J.; Jiao, L.; An, J. Sig-NMS-Based Faster R-CNN Combining Transfer Learning for Small Target Detection in VHR Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8534–8545. [CrossRef]
19. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]

20. Ding, J.; Xue, N.; Long, Y.; Xia, G.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
21. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Trans. Multimedia* **2018**, *20*, 3111–3122. [CrossRef]
22. Ding, P.; Zhang, Y.; Deng, W.J.; Jia, P.; Kuijper, A. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 208–218. [CrossRef]
23. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
24. Zhu, Y.; Du, J.; Wu, X. Adaptive Period Embedding for Representing Oriented Objects in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7247–7257. [CrossRef]
25. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2337–2348. [CrossRef]
26. Xu, Z.; Xu, X.; Wang, L.; Yang, R.; Pu, F. Deformable convnet with aspect ratio constrained nms for object detection in remote sensing imagery. *Remote Sens.* **2017**, *9*, 1312. [CrossRef]
27. Ren, Y.; Zhu, C.; Xiao, S. Deformable Faster R-CNN with aggregating multi-layer features for partially occluded object detection in optical remote sensing images. *Remote Sens.* **2018**, *10*, 1470. [CrossRef]
28. Wang, C.; Bai, X.; Wang, S.; Zhou, J.; Ren, P. Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 310–314. [CrossRef]
29. Chen, J.; Wan, L.; Zhu, J.; Xu, G.; Deng, M. Multi-Scale Spatial and Channel-wise Attention for Improving Object Detection in Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 681–685. [CrossRef]
30. Lu, X.; Zhang, Y.; Yuan, Y.; Feng, Y. Gated and Axis-Concentrated Localization Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 179–192. [CrossRef]
31. Pang, J.; Li, C.; Shi, J.; Xu, Z.; Feng, H. $\mathcal{R}^2$ -CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5512–5524. [CrossRef]
32. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
33. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
34. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
35. Tian, Z.; Wang, W.; Zhan, R.; He, Z.; Zhang, J.; Zhuang, Z. Cascaded Detection Framework Based on a Novel Backbone Network and Feature Fusion. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2019**, *12*, 3480–3491. [CrossRef]
36. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 734–750.
37. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636.
38. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
39. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
40. Liu, X.; Liang, D.; Yan, S.; Chen, D.; Qiao, Y.; Yan, J. FOTS: Fast Oriented Text Spotting With a Unified Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5676–5685.

41. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 8232–8241.
42. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the Computer Vision and Pattern Recognition. IEEE/CVF Conference 2018 (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
43. Sun, P.; Chen, G.; Shang, Y. Adaptive Saliency Biased Loss for Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7154–7165. [CrossRef]
44. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-Merged Single-Shot Detection for Multiscale Objects in Large-Scale Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3377–3390. [CrossRef]
45. Gong, Y.; Xiao, Z.; Tan, X.; Sui, H.; Xu, C.; Duan, H.; Li, D. Context-Aware Convolutional Neural Network for Object Detection in VHR Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 34–44. [CrossRef]
46. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Korner, M.; Reinartz, P. Towards Multi-class Object Detection in Unconstrained Remote Sensing Imagery. In Proceedings of the ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 4–6 December 2018; pp. 150–165.
47. Guo, H.; Yang, X.; Wang, N.; Song, B.; Gao, X. A Rotational Libra R-CNN Method for Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5772–5781. [CrossRef]
48. Liu, W.; Ma, L.; Wang, J.; Chen, H. Detection of Multiclass Objects in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 791–795. [CrossRef]
49. Li, Z.; Tang, X.; Wu, X.; Liu, J.; He, R. Progressively Refined Face Detection Through Semantics-Enriched Representation Learning. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 1394–1406. [CrossRef]
50. Tang, X.; Du, D.K.; He, Z.; Liu, J. PyramidBox: A Context-assisted Single Shot Face Detector. In Proceedings of the ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 797–813.
51. Lin, G.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 1–26 July 2017; pp. 1925–1934.
52. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
54. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
55. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *arXiv* **2019**, arXiv:1908.05612.