Article

FESSD: Feature Enhancement Single Shot MultiBox Detector Algorithm for Remote Sensing Image Target Detection

Jianxin Guo, Zhen Wang * and Shanwen Zhang

School of Electronic Information, Xijing University, 710123 Xi'an, China;
guojx_516@126.com (J.G); wjdw716@163.com (S.Z)
* Correspondence: miswz@iocas.ac.cn; Tel.: +86-132-7936-7151 (Z.W.)

Abstract: Automatic target detection of remote sensing images (RSI) plays an important role in military reconnaissance, disaster monitoring, and target rescue. The core task of remote sensing target detection is to judge the target categories and complete precise location. However, the existing target detection algorithms have limited accuracy and weak generalization capability for remote sensing images with complex backgrounds. To achieve accurate detection of different categories targets in remote sensing images, this study presents a novel feature enhancement single shot multibox detector (FESSD) algorithm for remote sensing target detection. The FESSD introduces feature enhancement module and attention mechanism into the convolution neural networks (CNN) model, which can effectively enhance the feature extraction ability and nonlinear relationship between different convolution features. Specifically, the feature enhancement module is used to extract the multi-scale feature information, and enhance the model nonlinear learning ability; the self-learning attention mechanism (SAM) is used to expand the convolution kernel local receptive field, which makes the model extract more valuable features. In addition, the nonlinear relationship between different convolution features is enhanced using the feature pyramid attention mechanism (PAM). The advantage of FESSD over other state-of-the-art target detection methods is validated by experiments on the presented seven-class target detection dataset (SD-RSI) and the public DIOR dataset.

Keywords: remote sensing image (RSI); target detection; convolution neural networks (CNN); FESSD; feature enhancement)

1. Introduction

Remote sensing image (RSI) target detection is one of the hot-research problems in the (RSI) interpretation field, which relates to many essential and fundamental applications in both military and civilian fields [1–3]. In the specific applications, remote sensing target detection can be applied in ship location [4], vehicle counting [5], disaster rescue [6], target attacks [7], warfare analysis [8], and so on. However, compared with the ordinary optical images, remote sensing images have the characteristics of complex background, high target density, small target size, and large feature similarity between different categories targets [9,10], which pose many challenges to the accurate detection and localization of different target in remote sensing images.

In recently years, many RSI target detection methods have been proposed and widely used in different fields. The existing RSI target detection methods can be classified into traditional machine learning (ML) methods and deep learning (DL)-based methods [11]. To achieve accurate target detection, the ML-based methods mainly include feature extraction and classifier design. In the feature extraction phase, many methods are used for target detection. The extracted features are specific to the target characteristics, such as color, texture, shape, angle, and so on [12]. The commonly used feature extraction methods include scale-invariant feature transform (SIFT) [13], histogram of oriented gradient (HOG) [14], and deformable part model (DPM) [15]. In addition, many multi-feature combination methods have been proposed for RSI target detection, which can effectively enhance the

 doi:10.20944/preprints202212.0186.v1

feature extraction effect [16–19]. The purpose of designing classifiers is to determine the specific categories of target, and the commonly used classifiers include support vector machine (SVM) [20], random forest (RF) [21], decision tree model (DTM) [22], and naive bayes classifier (NBC) [23]. However, both feature extraction and classifier selection is designed based on experience. Although better detection results can be achieved for specific application scenarios, it is highly dependent on prior knowledge, resulting in poor adaptability and generalization capabilities [24]. Recently, deep learning (DL) has achieved more significant breakthroughs and success in many application fields on image processing and computer vision [25–28]. The DL model can actively learn features from the data, which provides a new notion and framework for RSI target detection. DL-based target detection methods can be divided into regression-based single-stage detection methods [29], and region proposal-based two-stage target detection methods [30]. The single-stage target detection methods are based on the regression perspective, which regress the edge position and category of the target directly on the input image. The single-stage detection methods include DSSD [31], RetinaNet [32], RefineDet [33], and so on. The two-stage target detection methods generates a series of region proposal boxes on the original image, then input the region proposal boxes and feature maps to the region of interest pooling (ROI Pooling) layer, and finally the classifier is used to achieves target classification and localization. The currently proposed two-stage target detection methods include Faster R-CNN [34], Mask R-CNN [35], FPN [36], and Cascade RCNN [37]. The DL-based target detection methods use the convolution neural network (CNN) structure for image feature extraction [38], which presents encouraging performances on target detection. However, due to the significant difference between the RSI and optical images, so using the single convolution cannot fully extract the target region features.

In this study, to enhance the model feature extraction ability and suppress background information interference, we propose a feature enhancement single shot multibox detector algorithm (FESSD), where the feature enhancement module, self-attention mechanism (SAM), and feature pyramid attention mechanism (FPA) are combined with the original SSD algorithm. In FESSD, the feature enhancement module is used to enhance the shallow features and deep features extracted by different convolution layer, and the self-attention mechanism is used to expand the local and global receptive fields while enhancing the correlation between different features. Moreover, the feature pyramid attention mechanism is used for multi-scale feature fusion and improving the model nonlinear learning ability. In summary, the main contributions of this study are as follows,

- To improve the accuracy of remote sensing target detection, this study proposed a feature enhancement single shot multibox detector algorithm (FESSD). Different from previous DL-based methods, the FESSD can effectively extract multi-scale features information of remote sensing images.
- Feature enhancement structures that including feature enhancement module, self-attention mechanism, and feature pyramid attention mechanism are proposed for extracting meaningful features of remote sensing targets and suppressing background feature information interference.
- To assess the performance of the proposed FESSD, we perform a wide range of comparisons between different target detection methods on SD-RSI and DIOR dataset. The experiments results show that the proposed method is far more efficient than the other state-of-art methods.

2. Related Work

The related works are reviewed in this section, including the traditional machine learning (ML)-based methods and deep learning (DL)-based methods for target detection.

2.1. ML-Based Remote sensing Target Detection Method

In the past decades, many ML-based methods have been proposed for remote sensing target detection. The main steps of ML-based methods include feature extraction, feature

 doi:10.20944/preprints202212.0186.v1

selection, and category classification. Specifically, Dong et al. [39] extract multi-scale features of remote sensing images and uses random forest metric learning (RFML) as a classifier for remote sensing target detection. Li et al. [40] proposed an automatic target detection using contour spatial model, which uses dynamic programming to calculate the similarity between contour information and target templates to achieve target detection. In [41], the sparse representation and hough transform (HT) are combined for target detection, the learned target and background dictionaries are used to represent sparse images specific classes, and the hough voting is used for spatial feature integration. The method of multi-feature fusion is an effective way to achieve accurate target detection. In [42], multiple features including color, texture, shape, density, orientation, etc. are combined to realize target detection. Zhu et al. [43] presents a novel target detection method based on both bottom-up saliency and top-down saliency, where the scale-invariant features transform and SVM are used to detection the target regions. In [44], it combined with the target information and first-order Markov model to train the nonlinear support vector data description (SVDD) and conduct target classification. Yang et al. [45] propose a novel target detection framework sparse CEM and sparse ACE based on the constrained energy minimization (CEM) and the adaptive coherence estimator (ACE). Zhang et al. [46] proposed a regularization framework for the measurement matrices, which adds a scaled identity matrix to strengthen the inverse matrices stability and improve the detection accuracy. However, the ML-based methods are susceptible to interference from complex scenes and background information, thus limiting its robustness and generalization.

2.2. DL-Based Remote sensing Target Detection Method

Benefit from the powerful feature extraction ability of CNN, it has been widely used in remote sensing target detection. To enhance the feature extraction effect of CNN, Lei et al. [47] proposed a region-enhanced CNN (RECNN) for remote sensing images target detection. In RECNN, it uses saliency constraint and multilayer feature fusion strategy to enhance the CNN model detection performance. Lu et al. [48] proposed a new target detection model, which uses the channel attention mechanism to learn the global and local features of the target, and by axis-concentrated prediction (ACP) to project the convolution feature maps into different spatial directions. To obtain a suitable ROI scale, Dong et al. [49] carried out the statistical analysis on the target, and designed a better target detection model based on the suitable ROI scale. The design of loss function plays an important role in improving the model generalization ability and target detection accuracy. In [50], Sun et al. proposed a new method for designing loss function, called adaptive saliency biased loss (ASBL), which can train target detectors model to achieve better performance. Bai et al. [51] use the dense residual network (DRNet) and ROI pooling to enhance the original Faster R-CNN detection performance, so it can applicable to the field of remote sensing target detection. The YoLOv3 (You Only Look Once) is a commonly used single stage target detection method. In [52], Ma et al. use the ShuffleNet to enhance the feature extraction capability of YOLOv3 and use the generalized intersection over union (CIOU) loss improves the model detection performance. To alleviate the influence of illumination transformation and complex background on the detection performance for remote sensing target, Li et al. [53] proposed a novel remote sensing target detection algorithm, GLS-Net (Global-Local Saliency Constraint Network), which can make full use of global semantic feature information and achieve more accurate oriented bounding boxes. The use of multi-scale features contained in remote sensing images is important for improving the target detection accuracy, Sun et al. [54] proposed a part-based convolutional neural network (PBNet) for complex composite object detection in remote sensing imagery, which uses the context refinement module to obtain the multi-scale context features and global context features, effectively improving the remote sensing object detection accuracy. To improve the model feature extraction ability, He et al. [55] proposed a deformable contextual feature pyramid module for adaptive extraction the multi-scale features contained in remote sensing images, and the experimental results demonstrated the importance of extracting

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

multi-scale features for improving detection accuracy. Wang et al. [56] proposed a feature-merged single-shot detection (FMSSD) for multiscale objects detection in remote sensing images, which uses the atrous spatial feature pyramid (ASFP) module to extract and fuse the multi-scale features contained in remote sensing images, and the area-weighted loss function is used for improve the detection accuracy for small targets.

3. Proposed Method

3.1. Overall Architecture

SSD as a single-stage target detection algorithm has achieved better results in ordinary optical image detection [57]. However, limited by the ability to obtain deep features and global context information, SSD cannot achieve accurate detection of small and dense targets. To improve the detection accuracy of small and dense target in remote sensing images, the FESSD is proposed based on the SSD algorithm. In FESSD, the shallow feature enhancement (SFE) module, deep feature enhancement (DFE) module, self-attention mechanism (SAM), and feature pyramid attention mechanism (PAM) are introduced into SSD algorithm to improve the detection accuracy of remote sensing image targets. The overall framework of FESSD is shown in Fig. 1, where the SFE module and DFE module are used to perform shallow and deep feature enhancement; the SAM is introduced into the backbone network to enhance the correlation of the extracted features by different convolution layer; and the PAM is used to fully extract local and global feature information of target regions.

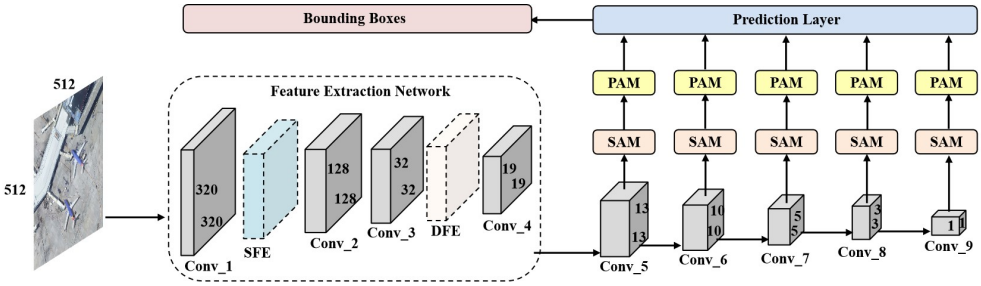

Figure 1. The structure of feature enhancement single shot multiBox detector (FESSD), including the **Shallow Feature Enhancement (SFE) Module**, **Deep Feature Enhancement (DFE) Module**, **Self-Attention Mechanism (SAM)**, and **Feature Pyramid Attention Mechanism (PAM)**.

3.2. Feature Enhancement Module

To enhance the model feature extraction ability and relationships between different features, inspired by Inception [58], ResNet [59], and dual-path network (DPN) [60], we proposed shallow feature enhancement (SFE) module and deep feature enhancement (DFE) module. The structure of feature enhancement module is shown in Fig. 2.

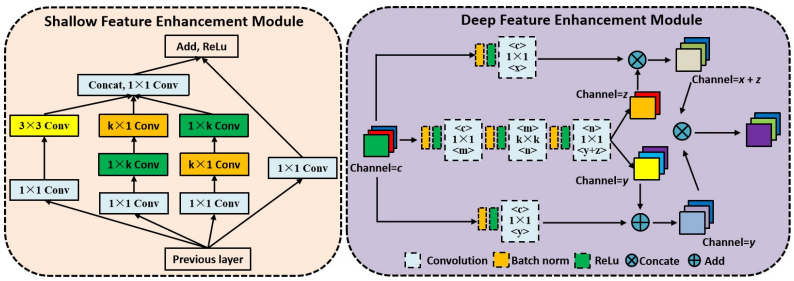

Figure 2. The structure of feature enhancement module, including the **Shallow Feature Enhancement (SFE) Module** and **Deep Feature Enhancement (DFE) Module.**

Shallow Feature Enhancement (SFE) Module. The SFE module contains four different branches, the first branch uses 1×1 convolution and 3×3 dilated convolution to enhance the model global receptive field range and nonlinear learning ability. the second and third branches uses group convolution to improve the feature extraction ability without

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

increasing calculation parameters; the fourth branch uses 1×1 convolution to obtain original primary features and uses residual connection for feature transfer. Specially, the SFE module uses group convolution to decompose the $k \times k$ convolution kernel into $1 \times k$ and $k \times 1$, which saves training time while ensure the local receptive field unchanged, and the *Concat* function and *Add* fusion operation are used to fuse the different branch feature maps. The formal description is as follows,
(1) The convolution layer is defined as follows,

$$x^l = f\left(W^l x^{l-1} + b^l\right) \tag{1}$$

where, l represents the number of layers, W represents the convolution weight, b represents the bias value, and f represents activation function.
(2) The definition of dilated convolution is as follows,

$$(F \circledast l)(p) = \sum_{s+lt} F(s)k(t) \tag{2}$$

where, $\circledast$ represents the dilated convolution, F represents the input image, k represents kernel function, and p , s , and t represent the corresponding domains, respectively.
(3) The calculation process of residual learning is as follows,

$$x_{l+1} = x_l + F(x_l, W_l) \tag{3}$$

where, x_l represents the current layer, x_{l+1} represents the next layer, W_l represents the weight of l layer, and F represents residual function.
(4) The calculation formula of *Concat* function and *Add* fusion operation are as follows,

$$Z_{concat} = \sum_{i=1}^{c} x_i * k_i + \sum_{i=1}^{c} y_i * k_{i+c} \tag{4}$$

$$Z_{add} = \sum_{i=1}^{c} (x_i + y_i) * k_i \tag{5}$$

where, x_i and y_i represent the input of different channels, c represents the number of channels, k represents the kernel function, and $*$ represents the convolution operation.
Deep Feature Enhancement (DFE) Module. The DFE module combined the advantages of ResNet [59] and DenseNet [61] for feature extraction, which can not only deepen the model structure, but also realize the progressive fusion of different layer feature information. The DFE module uses the group convolution and dense connection operation to enhance the feature information of different convolution layers and channels. The calculation process of DFF module is as follows, (1) The dense connection calculation formal is as follows,

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{6}$$

where, $H_l(\cdot)$ denotes the transformation function, and l denotes convolution layers.
(2) Group convolution. The input of group convolution is $H \times W \times C_1$, C_2 is its filter, and the output by convolution is $H \times W \times C_2$. Therefore, the number of calculation parameters for group convolution is half of the corresponding origibal convolution operation.
(3) The calculation formula of *Concat* functions and *Add* fusion as Eq. 4 and Eq. 5.

3.3. Self-Attention Mechanism

In the operation process of CNN, since the limited and fixed size of the convolution kernel, each convolution operation can only cover the area calculated by the convolution kernel, so the global feature and multi-scale feature contained in the image are not easily obtained, and ignores the correlation between different features. To make the features extracted by the convolution layer more relevant, the self-attention mechanism is introduced in FESSD, and the structure of self-attention is shown in Fig. 3. The self-attention mechanism (SAM) is mainly used to perform multi-scale transformation and fusion of input feature maps, to enhance the correlation between different feature maps and improve

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

the detection accuracy. The SAM module includes convolution, feature transformation, feature scaling, and feature fusion. The convolution operation can normalize the channel of input feature map; the feature transformation and feature scaling are used to enhance the correlation between different feature points in the feature map; and the feature fusion output the final self-attention feature map.

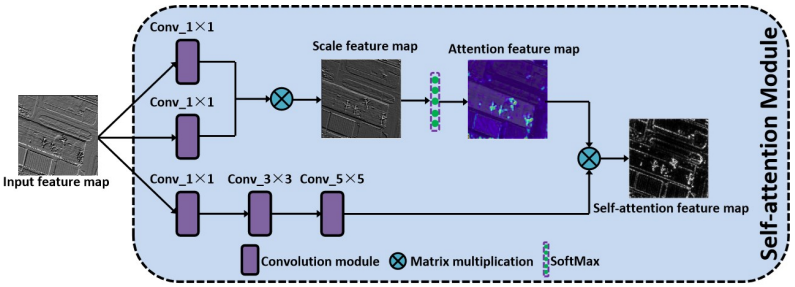

Figure 3. The structure of self-attention mechanism (SAM), including the **convolution module**, **matrix multiplication**, and **SoftMax** function.

As shown in Fig. 3, assuming that the size of input feature map N is $c \times w \times h$, where c represents the number of feature map channels, w represents the width, h represents the height. The overall calculation process of SAM is as follows,

(1) Perform three convolution operations on the input feature map, where the first and second convolution modules uses the 1×1 convolution kernel to compress the number of feature map channels, and then expand and transform the width and height of the feature map into matrix Q and K; the third convolution module continues to uses 1×1 convolution, but maintains the number of feature channels unchanged, and expands the width and height of the input feature map into matrix V; then the 3×3 and 5×5 convolution operations are performed to further extract the feature information.

(2) The matrix Q is transposed, and the transpose matrix Q^T and matrix K are multiplied to obtain the scale feature map matrix $E = Q^T K$.

(3) The scale feature matrix E is is normalized in the column direction using SoftMax function, and the relationship between different feature points of the feature map is obtained, that is the attention matrix feature map, and the matrix elements is calculated as follows,

$$\beta_{j,i} = \exp(E_{ij}) / \sum_i^N \exp(E_{ij}) \tag{7}$$

where, $\beta_{j,i}$ represents each element in the attention matrix feature map, E_{ij} represents each element in the scale feature map matrix, N represents the number of elements in the scale feature map matrix.

3.4. Feature Pyramid Attention Mechanism

The feature pyramid network (FPN) can extract feature of different scales from the pixel-level, and calculates the multiple receptive field information in parallel [36]. However, in the traditional feature pyramid network, the information fusion between of different scales feature maps is completed by simple linear superposition, ignoring the nonlinear relationship between of different levels branches. The feature pyramid attention mechanism (FPAM) uses the modified feature pyramid structure to obtain the multi-scale features of different objects, and increases the model nonlinear feature extraction ability.

As shown in Fig. 4, FPAM includes pyramid feature extraction (PFE) module and attention feature fusion (AFF) module, where the PFE module is used to obtain multi-scale features of different feature map; AFF module is used for multi-scale fusion of different scale attention feature maps. Specially, the PFE module includes four layers of convolution structure, in which 3×3, 5×5, 7×7, and 9×9 convolution operation are used for multi-scale feature extraction, and the down-sampling is used to reduce the feature map resolution size; the AFF module is constructed with a hierarchical manner, the multi-scale feature extraction is performed by using different size convolution kernels, and the up-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

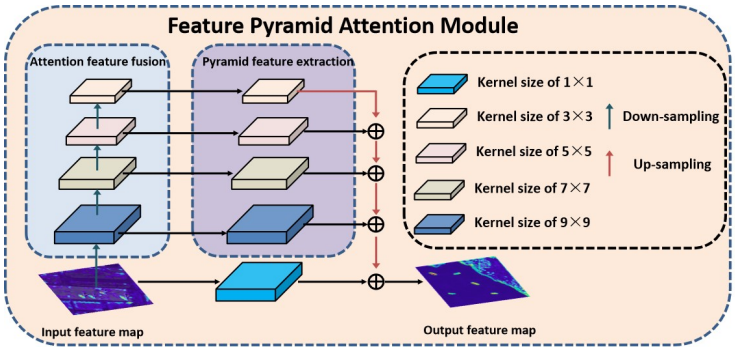

Figure 4. The structure of feature pyramid attention mechanism (FPAM), including the **Attention Feature Fusion (AFF) Module** and **Pyramid Feature Extraction (AFF) Module**.

sampling operation is used to generate the attention weight map. Moreover, to avoid the misjudgment of the original input feature map in the process of PFE and AFF module, the original feature map is linearly superimposed with the result of AFF module. The FPAM can mapping the feature into different sub-regions, it by aggregating the semantic content of different sub-regions to make the final output feature have global semantic information. The formal description of FPAM is as follows,

$$S = Conv_{1\times1}(I) + fpam(N) \tag{8}$$

where, S represents the output, I represents the input, and N represents the number of layers in the FPAM.

$$fpam(N) = Conv_{m\times n}(I_i) + U(fpam(i-1)) \tag{9}$$

where, $m \times n$ represents the size of convolution kernel, U represents up-sampling operation, $fpam(i-1)$ represents the output of the $(i-1)th$ layer.

$$I_i = Conv_{3\times3}(D(I_{i-1})) \tag{10}$$

where, D represents down-sampling operation, I_i represents the input of the ith layer.

3.5. Bounding Boxes Selection and Loss Function

To accurately detect remote sensing objects of different categories, the bounding boxes with different aspect ratios are designed to match the size of different object categories. Assuming that the convolution feature of m convolution layers are selected for the object detection, the size of candidate bounding boxes for the feature of ith layer is calculated as

$$S_i = S_{\min} + \frac{S_{\max} - S_{\min}}{m-1}(i-1)\ \ i \in [1, m] \tag{11}$$

where, $S_{\min}$ and $S_{\max}$ represent the scale coefficients of candidate bounding boxes.

Assuming that the ratio of width to height of the candidate bounding boxes is a_r, the width is $w_i = S_i \times \sqrt{a_r}$, and height is $h_i = S_i / \sqrt{a_r}$, the center coordinate is calculated as

$$(x, y) = \left(\frac{i+0.5}{|f_k|}, \frac{j+0.5}{|f_k|}\right)\ \ i, j \in [0, |f_k|] \tag{12}$$

where, $|f_k|$ represents the size of kth layer features.

To solve the problem of model degradation caused by the imbalance of the positive and negative samples of the remote sensing image during the model training process, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

FESSD is optimized and trained based on the SSD algorithm loss function combined with the focal classification loss, which is expressed as follows,

$$L(x,c,p,l,p) = \frac{1}{N}\left[L_{f1}(x,c,p) + aL_{loc}(x,l,g)\right] \tag{13}$$

where, N denotes the number of bounding boxes that match the ground truth, x denotes the input image, c is the object category, p is the predicted category probability, l denotes the bounding boxes, and a denotes the weight of bounding boxes and ground truth box.

The $L_{f1}(x,c,p)$ and $L_{loc}(x,l,g)$ are focus classification loss and bounding box regression loss, where the bounding box regression loss is inspired by the position regression function of Faster R-CNN, which is calculated as follows,

$$L_{loc}(x,l,g) = \sum_{i\in Pos}^{N}\sum_{m\in\{c_x,c_y,w,h\}} x_{ij}^{k} smooth_{L1}\left(l_i^m - g_j^m\right) \tag{14}$$

where, x_{ij}^{k} represents the comparison between ith candidate boxes and jth ground truth boxes for kth category, l_i^m represents the candidate box value, and g_i^m represents the ground truth box value. The cross entropy function is sued to calculate the loss of focus classification loss function $L_{f1}(x,c,p)$, which is calculated as follows,

$$L_{f1}(x,c,p) = -\sum_{i\in Pos}^{N} x_{ij}^{k} p\log\left(c_i^k\right) - \sum_{i\in Neg}(1-p)\log\left(c_i^0\right) \tag{15}$$

where, c_i^0 represents the predict category corrects probalility, and c_i^k is the probability calcualted by SoftMax function.

4. Experiments

4.1. Dataset

To evaluate the detection performance of the proposed FESSD algorithm, a total of 1,972 optical remote sensing images with a size of 500×375 pixels containing seven kinds of targets in different scenes are collected from Google Earth, and the spatial resolution varies from 0.5 to 2m. The dataset is named SD-RSI, and the original image and annotation

Table 1. The sample statistics of constructed dataset.

Dataset	Class	Image	Instance	small	Medium	Large
Train Dataset	Airplane	320	2382	1491	890	1
	Ship	400	1317	659	588	70
	Bridge	140	143	25	43	75
	Vehicle	140	956	842	114	0
	Storage tank (ST)	220	1735	884	845	6
	Baseball diamond (BD)	165	543	442	75	26
	Tennis court (TC)	95	482	318	95	69
Test Dataset	Airplane	101	970	471	398	1
	Ship	70	415	210	195	10
	Bridge	36	40	7	0	33
	Vehicle	49	318	295	23	0
	Storage tank (ST)	78	759	374	385	0
	Baseball diamond (BD)	74	158	212	42	4
	Tennis court (TC)	84	232	195	20	17

is available at https://drive.google.com/drive/rs-drive. The specific target categories include airplane, ship, bridge, vehicle, storage tank (ST), baseball diamond (BD), and tennis court (TC), which contain 10,550 instance samples, 7,558 targets are used as the train dataset, and 2,992 targets are used as test dataset. The train samples are separately rotated with the angle of $\varphi = \{10^{\circ}, 20^{\circ}, ..., 350^{\circ}\}$, which extend the number of train samples by 40 times. In addition, according to the criterion, the collected optical remote sensing images are divided into four classes (i.e. small target: area<32^2m; medium target: 32^2m<area<96^2m; Large target: area>96^2m) for split the target size, the target data distribution is shown in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

Tab. 1. Compared with the existing dataset NWPU VHR-10 [62] and AID [63] dataset, the constructed dataset mainly focuses on small size and medium size targets, and the target distribution is denser.

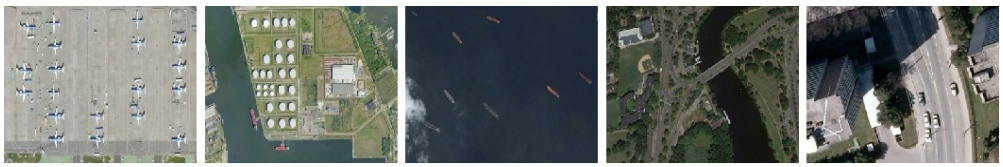

Figure 5. Samples of the constructed dataset.

Figure 6. Dataset image annotation process.

Fig. 5 shows sample examples of the constructed dataset, which are 19 airplanes, 29 storage tanks, 8 ships, 5 vehicles with a resolution of 1m, and 2 bridges with a resolution of 2m. Because some airplane and ship targets only contain a small number of pixels under the condition of low resolution, it is difficult to manually judge whether it is an airplane or ship. Therefore, only targets with more than 5 pixels are labeled. The images annotations process is shown in Fig. 6, in which the LabelImg software used for target annotation. In the obtained parameters, (x, y) represent the coordinates of the upper left corner in the rectangle box where the target is located, w denotes the width, and h denotes the height.

4.2. Evaluation Metrics

To evaluate the proposed MFENet, intersection over ratio (IoU), precision, recall, average precision (AP), and mean AP (mAP) are used to assess the performance of object detection in optical remote sensing images. The IoU is an important assessment indicator of whether the detection result is correct, which evaluates the degree of coincidence between the predicted bounding box D_{pre} and the true bounding box D_{gt} based on the Jaccard index. The IoU is defined as

$$IoU = \frac{\text{area}(D_{pre} \cap D_{gt})}{\text{area}(D_{pre} \cup D_{gt})} \tag{16}$$

where, $\text{area}(D_{pre} \cap D_{gt})$ represents the area of the intersection of D_{pre} and D_{gt}; $\text{area}(D_{pre} \cup D_{gt})$ is the area of union of D_{pre} and D_{gt}. In the object detection process, the overlap threshold ε_{iou} is given, if $IoU \geq \varepsilon_{iou}$, the predicted bounding box is considered to be true positive. The precision is defined as

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

where, TP represents the number of true positive predicted bounding box; FP represents the number of false positive predicted bounding box. The recall is defined as

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

where, FN represents the number of objects that are not detected. The area under the precision-recall curve (PRC) obtained by plotting precision and recall. F1_score is an equilibrium value of target detection accuracy, which is defined as

$$\text{F1\_score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{19}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

The AP is commonly used in object detection evaluation, which is defined as the average accuracy of object detection model at different recall levels. The mAP is the mean value of detection accuracy for different object categories, which is defined as

$$mAP = \sum_{n=0}^{N} AP_n / N \tag{20}$$

where, N represents the number of object categories.

4.3. Implementation Details

The experiments are conducted on a PC with 11GB NVIDIA 2080Ti, the operating system is Ubuntu 16.04, and the programming environment is Keras with CUDA kernels. The remote sensing image dataset is divided into train, validation, and test datasets. The train dataset includes 934 images, and each validation and test dataset include 310 remote sensing images. To prevent overfitting and obtain better generalization ability, based on the idea of transfer learning, the backbone network is pre-trained in the ImageNet dataset, and then the pre-trained model is trained in the remote sensing image dataset. The FESSD use the batch train method for experiment on the dataset, each batch contains 32 images. The model traverses all the train dataset is a single iteration, and the number of iterations is set to 10,000 in the experiment process. The objective function of FESSD is optimized by adaptive moment estimation (Adam) with the momentum of 0.9, and the batch size is set to 8 and weight decay of 0.005. The training process of FESSD is divided into three steps. The backbone network is pre-trained in the ImageNet dataset to the first stage. The initial learning rate of the first stage is set to 0.001, the decay coefficient is 0.1, and the decay is performed every 1,000 iterations. The second train phase is to train the pre-trained model for the remote sensing image dataset. In this phase, the transfer learning method is used to train the part weight of the pre-trained dataset. The second phase train parameter settings of the model are the same as the first phase. Finally, based on the model trained in the second phase, the overall model structure of FESSD is trained. The initial learning rate of this stage is set to 0.0001, and it decays 0.1 after 3,000 iterations.

4.4. Parameter Optimization

The IoU threshold is an important indicator that affects mAP. The higher thresholds of IoU, the more accurate the regression of the corresponding bounding boxes. Since the target detection in the remote sensing image is a two-class problem, that is, the image includes target area and backbone area. Therefore, when the predict value exceeds 0.5, it can be determined that the predict class is correct. Under the same experimental parameter settings, the IoU threshold is set to 0.2, 0.35, 0.5, 0.65 and 0.8. To select the appropriate IoU, the precision, recall, F1_score, and mAP indicators are calculated for different IoU thresholds on the test dataset, and the experimental results are shown in Tab.2. It can be seen from the experiment results, with the IoU threshold increases, the evaluation indicators of the model changed. The F1_score and mAP indicators first increase then decrease, and the optimal value is obtained when IoU reaches 0.5; From the Eq. (17), we can know that precision is determined by TP and FP, so with IoU increases, FP increases accordingly, which leads to the precision value decrease; From the Eq. (18), we can know that recall is determined by TP and FN, so with IoU increases, the TP value gradually increases, and the

Table 2. The impact of different IoU thresholds on model performance.

Method	IoU threshold	Precision	Recall	F1_score	mAP
FESSD	0.2	**95.18**%	79.82%	84.63%	79.25
	0.35	93.45%	81.26%	86.75%	81.26
	0.5	92.86%	82.51%	**88.45**%	**83.51**
	0.65	87.53%	83.14%	81.42%	80.65
	0.8	80.24%	**85.52**%	76.53%	78.32

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

recall also increases; Eq. (19) shows that the F1_score depends on precision and recall, so the change trend of F1_score is consistent with precision.

4.5. Performance Evaluation and Comparison

In this experiment, we evaluated the performance of FESSD algorithm, and compared it with multiple target detection algorithms, i.e., Contour-Based Spatial (CBS) Model [40], Partial Intensity Invariant Feature (PIIF) Descriptor [65], ASBL-RetinaNet [50], RECNN [47], GLS-Net [53], SSD [57], YOLOV3 [66], and Faster R-CNN [67]. The CBS model is a contour-based target detection method, which divides the image into multiple candidate regions, then uses dynamic programming to calculate the similarity between the contour information and the target template, and optimizes the detection result based on the contour similarity and the spatial relationship. The PIIE descriptor is a detection method based on hand-crafted feature, which used the UR-SHIFT to extract the multi-scale features, then uses the PIIE description to represent the multi-scale features, and the target detection is completed by the SVM classifier. The ASBL-RetinaNet proposed an adaptive saliency biased loss for training RetinaNet, which enhanced the model detection performance. The RECNN is a region-enhanced CNN for remote sensing images target detection, which introduces the saliency constraint and multilayer information fusion strategy into the CNN model. By combining a saliency algorithm with a feature pyramid network, the GLS-Net proposed a global-local saliency constraint network to reduce the impact of complex background. The SSD is a based on CNN one-stage target detection network with multi-scale feature extracted ability. The YOLOV3 is a speed and accurate target detection model, which can fully explore the spatial correlation of different feature maps. The Faster R-CNN is a two-stage target detection model, which combined region proposal network and object detection to complete end-to-end network training and testing. The results of nine target detection algorithms under the same experimental settings are shown in Tab. 3. Moreover, to evaluation the detection accuracy, the parameter IoU of all compared model is set to 0.5.

Table 3. Performance comparisons of FESSD and the other state-of-the-art methods.

Methods	Airplane	Vehicle	Ship	Bridge	ST	BD	TC	mAP	FPS
CBS-Model	0.697	0.518	0.658	0.689	0.795	0.763	0.803	0.703	5.3
PIIF	0.602	0.486	0.621	0.714	0.768	0.751	0.824	0.680	8.4
ASBL-RetinaNet	0.702	0.593	0.665	0.781	0.812	0.795	0.843	0.741	28.6
RECNN	0.757	0.635	0.645	0.771	0.835	0.821	0.828	0.770	21.5
GLS-Net	0.791	0.674	0.683	0.765	0.824	0.842	0.819	0.771	30.3
SSD	0.742	0.702	0.712	0.774	0.846	0.856	0.855	0.783	34.5
YOLOV3	0.758	0.683	0.724	0.769	0.793	0.803	0.798	0.761	31.8
Faster R-CNN	0.713	0.652	0.703	0.715	0.765	0.785	0.742	0.725	4.6
FESSD	**0.819**	**0.725**	**0.758**	**0.783**	**0.879**	**0.884**	**0.891**	**0.819**	**35.6**

Tab. 3 shows the detection performance of each algorithm for different target categories, and the evaluation indicators used include AP value, mAP, and FPS. Compared with the SSD algorithm, the mAP of FESSD is boosted from 0.783 to 0.819, which indicates the effectiveness of the proposed improved scheme. In terms of mAP, FESSD is better other compared state-of-the-art algorithms. Specifically, compared with the ML-based methods CBS-Model and PIIF Descriptor, the mAP is 0.116 and 0.139 higher respectively; compared with DL-based detection methods ASBL-RetinaNet, RECNN, and CLS-Net, the mAP of FESSD is 0.078, 0.049, and 0.048 higher respectively; compared with the one-stage detection method YOLOV3 and two-stage detection method Faster R-CNN, the mAP of the proposed method is 0.058 and 0.094 higher respectively. In terms of seven different target detection categories, the AP value of FESSD is better than other detection methods. Specifically, for small targets airplane, vehicle, ship, and storage tank, the AP value of FESSD outperforms the second-best by 0.028, 0.023, 0.034, and 0.033; for medium targets bridge, baseball diamond, and tennis court, the AP of FESSD outperforms the second-best by 0.002, 0.082, and 0.036. In addition, the FPS of FESSD reaches 35.6, which is better

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

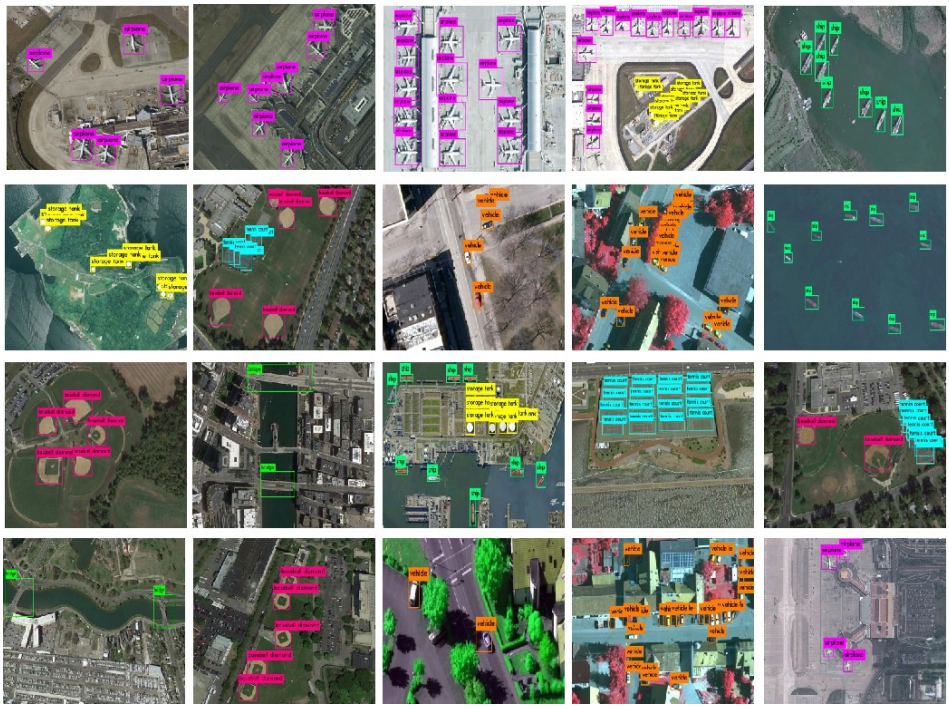

Figure 7. Visual detection results of FESSD on the SD-RSI dataset.

than other compared methods, indicating that FESSD has strong real-time performance. However, for small target of airplane, vehicle, and ship, although the FESSD achieved the best AP value, the results are not very satisfactory. The reason may be the small target has fewer pixels, the number of targets is dense, and has the target overlap phenomenon. To further demonstrate the effectiveness of FESSD for target detection, the visual detection results of different target classic are shown in Fig. 7. It can be observed from Fig. 7, FESSD has better detection results for multiple targets classic under different scenarios, even the remote sensing image with the large variations in orientations and sizes. Particularly, for small targets, the FESSD can achieve accurate detection and location.

Table 4. The performance comparison of different methods on the DIOR dataset.

Methods	Airplane	Harbor	Ship	Vehicle	BT	ST	mAP	FPS
CBS-Model	0.638	0.652	0.613	0.508	0.783	0.765	0.659	6.1
PIIF	0.584	0.597	0.568	0.486	0.752	0.735	0.620	8.3
ASBL-RetinaNet	0.671	0.698	0.602	0.552	0.823	0.782	0.688	29.4
RECNN	0.714	0.735	0.648	0.596	0.842	0.795	0.721	22.5
GLS-Net	0.738	0.759	0.668	0.624	0.865	0.813	0.744	31.6
SSD	0.674	0.682	0.617	0.542	0.776	0.724	0.669	35.8
YOLOV3	0.695	0.712	0.638	0.571	0.792	0.753	0.693	32.4
Faster R-CNN	0.684	0.673	0.625	0.542	0.758	0.712	0.665	5.3
FESSD	**0.812**	**0.793**	**0.782**	**0.758**	**0.862**	**0.864**	**0.812**	**36.5**

To further verify the performance of FESSD, we conducted experimental analysis on the DIOR dataset. The DIOR dataset contains more than 23k images and 192k instances, covering 20 object categories [68]. In the experimental process, we selected small and dense target categories such as airplane, harbor, ship, vehicle, baseball field (BF), and storage tank (ST) as detection objects, and the performance quantitative analysis of different methods is shown in Tab. 4. It can be seen from Tab. 4 that the mAP of FESSD on the DIOR dataset reaches 0.812, which is higher than other compared methods. The mAP of ML-based methods CBS-Model and PIIF are 0.659 and 0.620, indicating that these methods cannot accurately detect small and dense targets. DL-Based remote sensing target detection methods ASBL-RetinaNet, RECNN, and GLS-Net achieve better detection results, but the detection accuracy is relatively poor for small targets such as airplane, ship, and vehicle. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

Figure 8. Visual detection results of FESSD on the DIOR dataset.

mAP of generic target detection methods SSD, YOLOV3, and Faster R-CNN are 0.669, 0.693, and 0.665, although better than the ML-based methods, but cannot satisfy the requirements for accurate detection of remote sensing targets. In term of detection speed, the FPS of FESSD reaches 36.5, indicating that it can achieve fast detection of remote sensing targets. Fig. 8 shows the visual detection results of FESSD on the DIOR dataset, from which it can be seen that FESSD can achieve accurate detection and location of different remote sensing target categories.

4.6. Ablation Study

In this experiment, we evaluate the performance of each part of FESSD. To analyze the impact of the shallow feature enhancement (SFE) module and the deep feature enhancement (DEF) module on the remote sensing target detection accuracy mAP, the experiment set up different feature enhancement module combinations, and the experimental effects of each module are shown in Tab.5. It can be seen from Tab. 5 that by adding the SFE module

Table 5. The influence of feature enhancement module on target detection.

SSD	SFE	DEF	Class							mAP	FPS
			Airplane	Vehicle	Ship	Bridge	ST	BD	TC		
△			0.738	0.687	0.695	0.792	0.824	0.831	0.829	0.771	36.8
△	△		0.775	0.713	0.748	0.825	0.786	0.875	0.873	0.798	28.6
△		△	0.756	0.695	0.724	0.803	0.765	0.852	0.857	0.778	38.2
△	△	△	0.824	0.731	0.763	0.812	0.892	0.897	0.912	0.830	36.2

on the basis of SSD model, the mAP increased from 0.771 to 0.798, and the mAP of the small targets airplane, vehicle, and ship increased by 0.037, 0.026, and 0.053, respectively. However, due to the addition of SFE module, the model parameters increased and the FPS decreased by 8.2. When the DFE module is added on the basis of SSD, the target detection accuracy of each class is improved, and the DFE module makes the deep network structure more efficient, so that the FPS reaches 38.2. When the SFE module and DFE module are added at the same time, in terms of detection accuracy, the mAP of FESSD has reached 0.830, which is an improvement of 0.059 compared to SSD, and the detection accuracy of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

different class has been improved, especially the small target airplane, vehicle, and ship reached 0.824, 0.731, and 0.763 respectively.

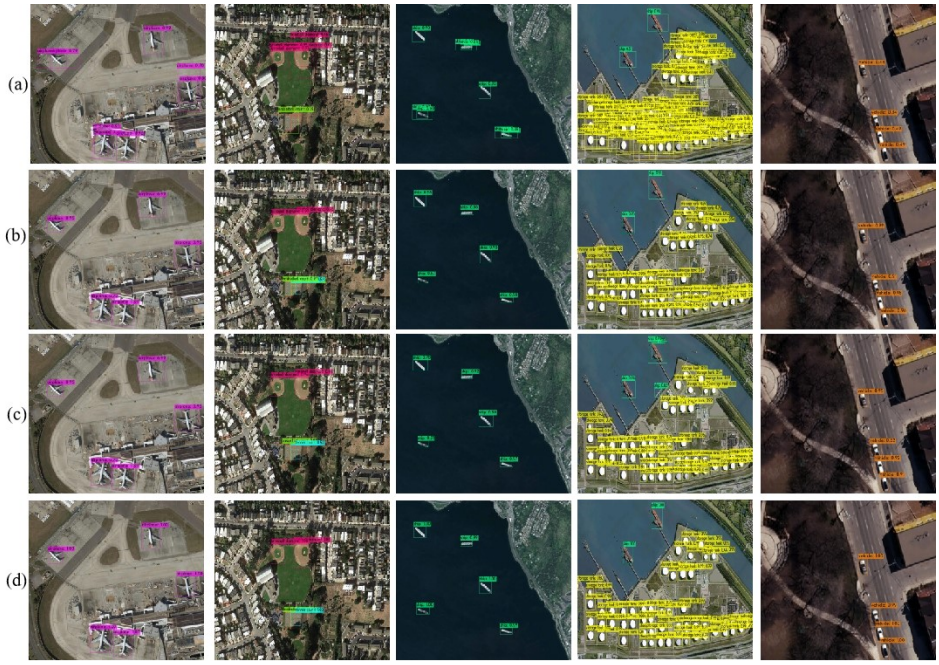

Figure 9. Visual detection results of different feature enhancement module.

Fig. 9 shows the visual detection results of different feature enhancement modules. Fig. 9 (a) is the detection result of SSD algorithm, which has problems such as target positioning offset and low confidence; Fig. 9 (b) is the detection results of the introduction SFE module, which has been greatly improved compared to the detection confidence of SSD; Fig. 9 (c) is the detection result of the introduction DEF module, although the confidence of target detection is improved, but occurs target positioning offset problem; Fig. 9 (d) is the detection result of FESSD, which not only accurately detect the target class, but also completes precise target positioning.

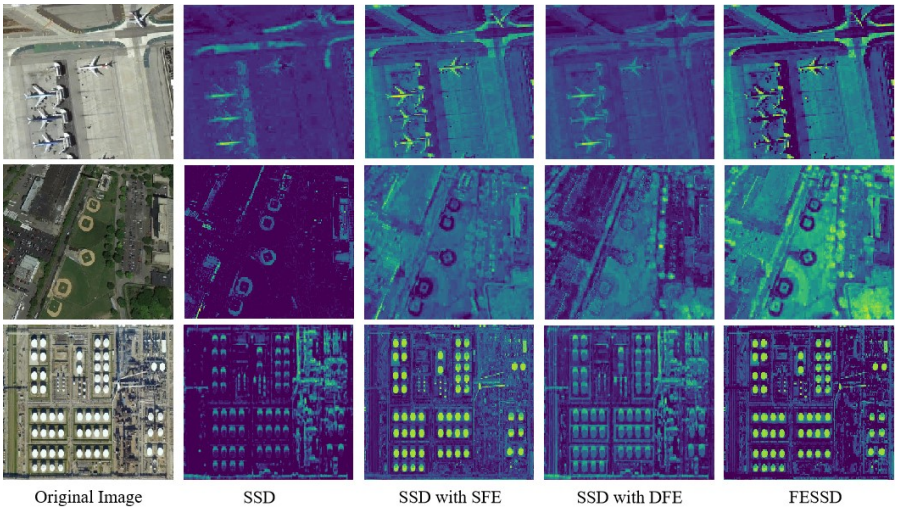

Figure 10. The feature maps visualization results of the last convolution layer.

To compare the feature enhancement effects of different feature enhancement modules on the network, we visualize the last convolution layer of FESSD, and the results are shown in Fig.10. It can be seen from the Fig. 10 that the original SSD algorithm has no obvious effect on the feature extraction of the detected target, and a large amount of target feature information is lost; When the SFE module is introduced, the feature extraction ability for target has been significantly enhanced, especially for the contour and texture features of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

small targets; When the DFE module is introduced, its feature extraction ability has been greatly improved compared to SSD, but the feature enhancement effect is not very obvious, because the function of DEF module is to deepen the depth of the network and alleviate the overfitting problem during the training process; FESSD has significant target area feature enhancement capabilities, it can be seen that FESSD accurately enhances the target area features and suppresses a large amount of background information.

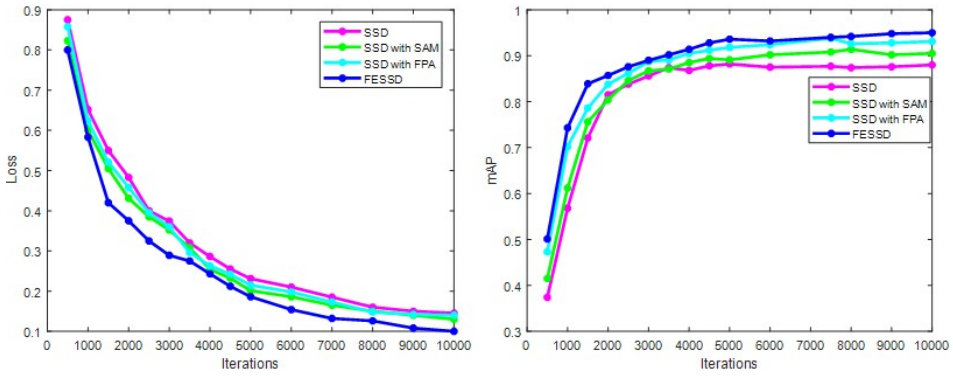

Figure 11. The effect of different attention mechanism on model performance.

The influence of different attention mechanisms on the model training process and detection accuracy is shown in Fig. 11. After iterative training, the loss value of SSD algorithm is 0.145, and the detection accuracy mAP reaches 0.832, indicating that its training effect and detection performance still need to be improved. When the attention mechanism module is introduced, both the SAM module and the FPA module improve the performance of the SSD algorithm. When the SAM module is introduced, the loss value and mAP of the model reaches 0.130 and 0.905. When the FPA module is introduced, the loss value and mAP of the model reaches 0.140 and 0.931. Compared with the SSD algorithm, the model with FPA and SAM has obvious advantages, and from the loss curve and mAP value, it can be analyzed that the FESSD model training effect and detection performance are better.

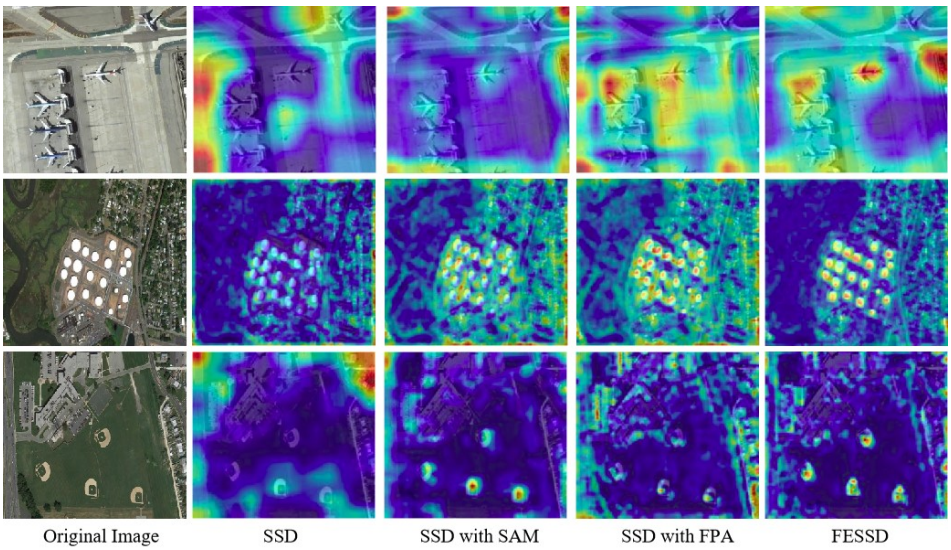

Figure 12. The effect of different attention mechanism on model performance.

Fig. 12 shows the heat maps of SSD algorithm, self-attention mechanism (SAM) module, feature pyramid attention mechanism (PAM) module, and FESSD algorithm. It can be seen from Fig. 12 that the region of interest (ROI) of the SSD algorithm does not correspond to the target area, which leads to inaccurate positioning of the target; the introduction of the SAM module enhances the correlation between features, so that ROI focuses on the target area, but there is still attention distraction phenomenon; the introduction of FPA enhances the feature fusion capability of the model, so that the ROI

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0186.v1

and the target area correspond to each other, and local receptive field is enlarged; When the SAM and FPA modules are introduced at the same time, the local receptive field and ROI of the model are expanded, and the ROI completely corresponds to the target area, which means that the model can accurately extract the features of the target area.

Table 6. The calculation efficiency comparison of different methods.

Methods	Memory Space (GB)	Train Time (h)	Test Time (s)	Parameters
ASBL-RetinaNet	10.72	12.15	6.74	48,175,462
RECNN	9.85	11.63	5.38	57,186,273
GLS-Net	9.26	13.54	5.75	63,241,156
SSD	8.53	9.83	4.26	42,381,268
YOLOV3	7.64	10.75	4.18	38,183,347
Faster R-CNN	10.56	14.68	7.83	62,248,751
FESSD	**6.38**	**7.25**	**2.41**	**32,168,285**

The operational efficiency of different DL-based methods is compared using the evaluation index of memory space (the memory utilization space of the model training process), calculation parameters, train time (the train time required to achieve the model optimal detection accuracy), and single image test time. The calculation efficiency comparison results of different methods are shown in Tab. 6. The proposed FESSD has advantages in different evaluation indicators of calculation efficiency, its memory space and calculation parameters are 6.38GB and 32,168,285, which show that FESSD requires less computation resource in the model training process. The train time of FESSD is 7.25h, which is the minimum of several compared methods, and its single image test time is 2.41s, further proves that FESSD can satisfy the real-time requirement of remote sensing target detection.

5. Conclusion

In this study, we presented a novel remote sensing target detection algorithm FESSD. The proposed method can effectively extract the multi-scale features and global contextual features contained in remote sensing images to achieve accurate detection of small and dense remote sensing targets. In FESSD, the feature enhancement module is used to enhance the shallow and deep features extracted by different convolution layers. The self-learning attention mechanism is used to expand the local receptive fields and multi-scale feature extract ability. In addition, the feature pyramid attention mechanism is used to enhance the nonlinear relationship between different feature maps. The experimental results on SD-RSI and DIOR show that proposed FESSD outperformed the compared state-of-the-art remote sensing target detection methods, which demonstrated the effectiveness and robustness of FESSD. In the future, we will introduce unsupervised learning methods to reduce the dependence on annotation data.

Author Contributions: Conceptualization, J.G. and S.Z.; methodology, J.G.; software, Z.W.; validation, J.G., Z.W. and S.Z.; formal analysis, Z.W. and S.Z.; writing—original draft preparation, J.G., Z.W. and S.Z.; writing—review and editing, J.G. and Z.W.; visualization, Z.W. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Neural Science Foundation of Shanxi Province under Grant 2021JQ-879, in part by the National Natural Science Foundation of China under Grant 62172338, in part by the National Natural Science Foundation of China under Grant 61671465, and in part by the Shaanxi Key Laboratory of Integrated and Intelligent Navigation open fund under Grant SKLIIN-20180211.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Zhengyang Zhao and Nan Xu for their support, secondly, thanks to Liping Wang and Feng Wang for their support.

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

References

1. Xiao, Z.; Gong, Y.; Long, Y.; Li, D.; Wang, X.; Liu, H. Airport Detection Based on a Multiscale Fusion Feature for Optical Remote Sensing Images. *IEEE Geosci. Remote Sensing Lett.* **2017**, *14*, 1469–1473.
2. Zhuang, Y.; Li, L.; Chen, H. Small Sample Set Inshore Ship Detection From VHR Optical Remote Sensing Images Based on Structured Sparse Representation. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing.* **2020**, *13*, 2145–2160.
3. Zhang, L.; Zhang, Y. Airport Detection and Aircraft Recognition Based on Two-Layer Saliency Model in High Spatial Resolution Remote-Sensing Images. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing.* **2017**, *10*, 1511–1524.
4. Li, Q.; Mou, L.; Liu, Q.; Wang, Y.; Zhu, X. X. HSF-Net: Multiscale Deep Feature Embedding for Ship Detection in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sensing.* **2018**, *56*, 7147–7161.
5. Zhou, H.; Wei, L.; Lim, C. P.; creighton, douglas; Nahavandi, S. Robust Vehicle Detection in Aerial Images Using Bag-of-Words and Orientation Aware Scanning. *IEEE Trans. Geosci. Remote Sensing.* **2018**, *56*, 7074–7085.
6. Thomas, J.; Kareem, A.; Bowyer, K. W. Automated Poststorm Damage Classification of Low-Rise Building Roofing Systems Using High-Resolution Aerial Imagery. *IEEE Trans. Geosci. Remote Sensing.* **2014**, *52*, 3851–3861.
7. Dai, Y.; Shen, L.; Cao, Y.; Lei, T.; Qiao, W. Detection of Vegetation Areas Attacked By Pests and Diseases Based on Adaptively Weighted Enhanced Global and Local Deep Features. In IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium; IEEE: Yokohama, Japan, 2019; pp. 6495–6498.
8. Zwieback, S.; Hajnsek, I.; Edwards-Smith, A.; Morrison, K. Depth-Resolved Backscatter and Differential Interferometric Radar Imaging of Soil Moisture Profiles: Observations and Models of Subsurface Volume Scattering. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing.* **2017**, *10*, 3281–3296.
9. Wang, Z.; Du, L.; Zhang, P.; Li, L.; Wang, F.; Xu, S.; Su, H. Visual Attention-Based Target Detection and Discrimination for High-Resolution SAR Images in Complex Scenes. *IEEE Trans. Geosci. Remote Sensing.* **2018**, *56*, 1855–1872.
10. Song, Z.; Sui, H.; Hua, L. How to Quickly Find the Object of Interest in Large Scale Remote Sensing Images. In IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium; IEEE: Valencia, 2018; pp. 4843–4845.
11. Pang, G.; Shen, C.; Cao, L.; Hengel, A. van den. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* **2022**, *54*, 1–38.
12. Zhu, H.; Ni, H.; Liu, S.; Xu, G.; Deng, L. TNLRS: Target-Aware Non-Local Low-Rank Modeling With Saliency Filtering Regularization for Infrared Small Target Detection. *IEEE Trans. on Image Process.* **2020**, *29*, 9546–9558.
13. Burger, W.; Burge, M. J. Scale-Invariant Feature Transform (SIFT). In Digital Image Processing; Texts in Computer Science; Springer International Publishing: Cham, 2022; pp. 709–763.
14. Guedira, M. R.; Qadi, A. E.; Lrit, M. R.; Hassouni, M. E. A Novel Method for Image Categorization Based on Histogram Oriented Gradient and Support Vector Machine. In 2017 International Conference on Electrical and Information Technologies (ICEIT); IEEE: Rabat, 2017; pp. 1–5.
15. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A Discriminatively Trained, Multiscale, Deformable Part Model. In 2008 IEEE Conference on Computer Vision and Pattern Recognition; IEEE: Anchorage, AK, USA, 2008; pp. 1–8.
16. Tokarczyk, P.; Wegner, J. D.; Walk, S.; Schindler, K. Features, Color Spaces, and Boosting: New Insights on Semantic Classification of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sensing.* **2015**, *53*, 280–295.
17. Zhao, B.; Zhong, Y.; Xia, G.-S.; Zhang, L. Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sensing.* **2016**, *54*, 2108–2123.
18. Peng, F.; Wang, L.; Gong, J.; Wu, H. Development of a Framework for Stereo Image Retrieval With Both Height and Planar Features. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing.* **2015**, *8*, 800–815.
19. Mylonas, S. K.; Stavrakoudis, D. G.; Theocharis, J. B.; Zalidis, G. C.; Gitas, I. Z. A Local Search-Based GeneSIS Algorithm for the Segmentation and Classification of Remote-Sensing Images. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing.* **2016**, *9*, 1470–1492.
20. Guo, Y.; Jia, X.; Paull, D. A Domain-Transfer Support Vector Machine for Multi-Temporal Remote Sensing Imagery Classification. In 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS); IEEE: Fort Worth, TX, 2017; pp. 2215–2218.
21. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing.* **2020**, *13*, 6308–6325.
22. Baraldi, A. Fuzzification of a Crisp Near-Real-Time Operational Automatic Spectral-Rule-Based Decision-Tree Preliminary Classifier of Multisource Multispectral Remotely Sensed Images. *IEEE Trans. Geosci. Remote Sensing* **2011**, *49*, 2113–2134.
23. Yang, J.; Ye, Z.; Zhang, X.; Liu, W.; Jin, H. Attribute Weighted Naive Bayes for Remote Sensing Image Classification Based on Cuckoo Search Algorithm. In 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC); IEEE: Shenzhen, 2017; pp. 169–174.
24. Zhang, L.; Zhang, L.; Tao, D.; Huang, X. Sparse Transfer Manifold Embedding for Hyperspectral Target Detection. *IEEE Trans. Geosci. Remote Sensing.* **2014**, *52*, 1030–1043.
25. Li, M.; Liu, Y.; Liu, X.; Sun, Q.; You, X.; Yang, H.; Luan, Z.; Gan, L.; Yang, G.; Qian, D. The Deep Learning Compiler: A Comprehensive Survey. *IEEE Trans. Parallel Distrib. Syst.* **2021**, *32*, 708–727.
26. Saha, S.; Mou, L.; Qiu, C.; Zhu, X. X.; Bovolo, F.; Bruzzone, L. Unsupervised Deep Joint Segmentation of Multitemporal High-Resolution Images. *IEEE Trans. Geosci. Remote Sensing.* **2020**, *58*, 8780–8792.

27. Li, G.; Li, L.; Zhu, H.; Liu, X.; Jiao, L. Adaptive Multiscale Deep Fusion Residual Network for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sensing.* **2019**, *57*, 8506–8521.
28. Alam, F. I.; Zhou, J.; Liew, A. W.-C.; Jia, X.; Chanussot, J.; Gao, Y. Conditional Random Field and Deep Feature Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sensing* **2019**, *57*, 1612–1628.
29. Hou, L.; Xue, J.; Lu, K.; Hao, L.; Rahman, M. M. A Single-Stage Multi-Class Object Detection Method for Remote Sensing Images. In 2019 IEEE Visual Communications and Image Processing (VCIP); IEEE: Sydney, Australia, 2019; pp. 1-4.
30. Chen, X.; Yu, X.; Ding, H.; Xue, Y.; Guan, J. Fast and Reftned Radar Processing For Maneuvering Target via Two-Stage Integration Detection. In 2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT); IEEE: Harbin, China, 2019; pp. 547–551.
31. Fu, C. Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A. C. DSSD: Deconvolutional Single Shot Detector. *arXiv* **2017**, arxiv:1701.06659
32. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327.
33. Zhang, S.; Wen, L.; Lei, Z.; Li, S. Z. RefineDet++: Single-Shot Refinement Neural Network for Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 674–687.
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149.
35. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In 2017 IEEE International Conference on Computer Vision (ICCV); IEEE: Venice, 2017; pp. 2980–2988.
36. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Honolulu, HI, 2017; pp. 936–944.
37. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1483–1498.
38. Shao, L.; Wu, D.; Li, X. Learning Deep and Wide: A Spectral Method for Learning Deep Networks. *IEEE Trans. Neural Netw. Learning Syst.* **2014**, *25*, 2303–2308.
39. Dong, Y.; Du, B.; Zhang, L. Target Detection Based on Random Forest Metric Learning. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing.* **2015**, *8*, 1830–1838.
40. Yu Li; Xian Sun; Hongqi Wang; Hao Sun; Xiangjuan Li. Automatic Target Detection in High-Resolution Remote Sensing Images Using a Contour-Based Spatial Model. *IEEE Geosci. Remote Sensing Lett.* **2012**, *9*, 886–890.
41. Yokoya, N.; Iwasaki, A. Object Detection Based on Sparse Representation and Hough Voting for Optical Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing.* **2015**, *8*, 2053–2062.
42. Dingwen Zhang; Junwei Han; Gong Cheng; Zhenbao Liu; Shuhui Bu; Lei Guo. Weakly Supervised Learning for Target Detection in Remote Sensing Images. *IEEE Geosci. Remote Sensing Lett. 2015*, **12**, 701–705.
43. Dan Zhu; Bin Wang; Liming Zhang. Airport Target Detection in Remote Sensing Images: A New Method Based on Two-Way Saliency. *IEEE Geosci. Remote Sensing Lett.* **2015**, *12*, 1096–1100.
44. Sakla, W.; Chan, A.; Ji, J.; Sakla, A. An SVDD-Based Algorithm for Target Detection in Hyperspectral Imagery. *IEEE Geosci. Remote Sensing Lett.* **2011**, *8*, 384–388.
45. Shuo Yang; Zhenwei Shi. SparseCEM and SparseACE for Hyperspectral Image Target Detection. *IEEE Geosci. Remote Sensing Lett.* **2014**, *11*, 2135–2139.
46. Zhang, Y.; Du, B.; Zhang, L. A Sparse Representation-Based Binary Hypothesis Model for Target Detection in Hyperspectral Images. *IEEE Trans. Geosci. Remote Sensing.* **2015**, *53*, 1346–1354.
47. Lei, J.; Luo, X.; Fang, L.; Wang, M.; Gu, Y. Region-Enhanced Convolutional Neural Network for Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sensing.* **2020**, *58*, 5693–5702.
48. Lu, X.; Zhang, Y.; Yuan, Y.; Feng, Y. Gated and Axis-Concentrated Localization Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sensing.* **2020**, *58*, 179–192.
49. 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 Sensing.* **2020**, *58*, 2104–2114.
50. Sun, P.; Chen, G.; Shang, Y. Adaptive Saliency Biased Loss for Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sensing.* **2020**, *58*, 7154–7165.
51. Bai, T.; Pang, Y.; Wang, J.; Han, K.; Luo, J.; Wang, H.; Lin, J.; Wu, J.; Zhang, H. An Optimized Faster R-CNN Method Based on DRNet and RoI Align for Building Detection in Remote Sensing Images. *Remote Sensing.* **2020**, *12*, 762–768.
52. Ma, H.; Liu, Y.; Ren, Y.; Yu, J. Detection of Collapsed Buildings in Post-Earthquake Remote Sensing Images Based on the Improved YOLOv3. *Remote Sensing.* **2019**, *12*, 44–51.
53. Li, C.; Luo, B.; Hong, H.; Su, X.; Wang, Y.; Liu, J.; Wang, C.; Zhang, J.; Wei, L. Object Detection Based on Global-Local Saliency Constraint in Aerial Images. *Remote Sensing.* **2020**, *12*, 1435–1442.
54. Sun, X.; Wang, P.; Wang, C.; Liu, Y.; Fu, K. PBNet: Part-Based Convolutional Neural Network for Complex Composite Object Detection in Remote Sensing Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing.* **2021**, *173*, 50–65.
55. He, Q.; Sun, X.; Yan, Z.; Fu, K. DABNet: Deformable Contextual and Boundary-Weighted Network for Cloud Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sensing.* **2022**, *60*, 1–16.

56. 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 Sensing*. **2020**, *58*, 3377–3390.
57. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A. C. SSD: Single Shot MultiBox Detector. *arxiv* **2016**, arXiv:1512.02325.
58. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Las Vegas, NV, USA, 2016; pp. 2818–2826.
59. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Las Vegas, NV, USA, 2016; pp. 770–778.
60. Chen, Y.; Li, J.; Xiao, H.; Jin, X.; Yan, S.; Feng, J. Dual Path Networks. *arxiv* **2017**, arXiv:1707.01629.
61. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. Densely Connected Convolutional Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Honolulu, HI, 2017; pp. 2261–2269.
62. 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 Sensing.* **2016**, *54*, 7405–7415.
63. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sensing*. **2015**, *7*, 14680–14707.
64. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. Densely Connected Convolutional Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Honolulu, HI, 2017; pp. 2261–2269.
65. Chen, S.; Zhong, S.; Xue, B.; Li, X.; Zhao, L.; Chang, C. I. Iterative Scale-Invariant Feature Transform for Remote Sensing Image Registration. *IEEE Trans. Geosci. Remote Sensing*. **2021**, *59*, 3244–3265.
66. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Las Vegas, NV, USA, 2016; pp. 779–788.
67. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149.
68. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*. **2020**, *159*, 296–307.