EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Transformer-Based Object Detection with Deep Feature Fusion Using Carafe Operator (TRCNet) in Remote Sensing Image

Shenao Chen[1], Bingqi Wang[2], Chaoliang Zhong[1*]

[1]Hangzhou Dianzi University, Hangzhou, China
[2]College of Science, Beijing Forestry University, Beijing, China

Abstract

Abstract: Recently, broad applications can be found in optical remote sensing images (ORSI), such as in urban planning, military mapping, field survey, and so on. Target detection is one of its important applications. In the past few years, with the wings of deep learning, the target detection algorithm based on CNN has harvested a breakthrough. However, due to the different directions and target sizes in ORSI, it will lead to poor performance if the target detection algorithm for ordinary optical images is directly applied. Therefore, how to improve the performance of the object detection model on ORSI is thorny. Aiming at solving the above problems, premised on the one-stage target detection model-RetinaNet, this paper proposes a new network structure with more efficiency and accuracy, that is, a Transformer-Based Network with Deep Feature Fusion Using Carafe Operator (TRCNet). Firstly, a PVT2 structure based on the transformer is adopted in the backbone and we apply a multi-head attention mechanism to obtain global information in optical images with complex backgrounds. Meanwhile, the depth is increased to better extract features. Secondly, we introduce the carafe operator into the FPN structure of the neck to integrate the high-level semantics with the low-level ones more efficiently to further improve its target detection performance. Experiments on our well-known public NWPU-VHR-10 and RSOD show that mAP increases by 8.4% and 1.7% respectively. Comparison with other advanced networks also witnesses that our proposed network is effective and advanced.

Keywords—Remote sensing image, transformer, target detection

Received on 30 May 2023, accepted on 21 August 2023, published on 23 August 2023

doi: 10.4108/ew.3404

1. Introduction

Over the past decades, up against the in-depth evolution of optical remote sensing technology, ORSI has owned better resolution. ORSI contain much more information. The object detection of ORSI is targeted at identifying high-value objects (aircraft, buildings, oil tanks, etc.) and locating them accurately, which has been broadly applied in urban planning [1] [2], military reconnaissance [3], etc.

In the wake of the deep learning framework evolution, innovations have found continual expressions in CNN-based target detection algorithms in the past ten years, with two important branches emerging as follows. The two-stage detection model is represented by RCNN and the single-stage one by yolo [4] [5]. After feature extraction using CNN, the two-stage detection model first uses RPN to generate high-quality RoI, then pools the RoI before finally regressing and classifying the bounding box. In contrast, the single-stage detection model directly regresses and classifies the bounding box. The two-stage model is slow and more accurate in the application. The single-stage model can rapidly function and achieve the real-time detection, but its accuracy is slightly defective. Therefore, this paper lays emphasis on the accuracy perfection of the single-stage target detection model while retaining its advantages.

[1]Corresponding author. Email: chaoliang_zhong@outlook.com

The CNN-premised target detection algorithm can achieve good results on ordinary optical images with simple and clear scenes, but many differences exist between ORSI and ordinary optical images taken by mobile phones [6]. The shooting of ORSI is done by satellites or aircraft flying at high altitudes. Long-distance shooting leads to the characteristics of multi-size, multi-resolution, and multi-direction. In addition, the background of the target is more complex with more diverse changes in background [7]. Traditional CNN has a limited receptive field, hindering the global information acquisition in the target recognition task of ORSI. Using stacking depth and pooling operation, the receptive field of CNN can be expanded. However, this will give rise to the degradation of small target detection performance.

Furthermore, FPN greatly promotes the development of a multi-scale target detection algorithm, which transmits high-level semantics, fuses it with low-level semantics after up-sampling to generate high-resolution and strong semantic feature maps, and enhances the detection performance of small targets. However, given that it adopts nearest neighbor up-sampling without incorporating the semantics of the feature map, it cannot effectively use semantics in feature fusion and reorganization [8].

This paper proposes TRCNet based on Retinanet, a single-stage detection model to tackle the aforementioned problems. As for the backbone, we use PVTv2 premised on the transformer to obtain the global information and do global modeling to eliminate the performance degradation of small target detection caused by insufficient receptive field and complex background of CNN. As for the Neck, we introduce the carafe operator to the FPN up-sampling process and guide the up-sampling process for efficient multi-features fusion. More specifically, the followings are our main contributions:

A new network structure, TRCNet, is blueprinted to detect multi-scale objects in ORSI with higher accuracy.

The network is perfected premised on Retinanet. In terms of the backbone, we introduce a transformer module for features extraction. With regard to the Neck, we utilized FPNcarafe to fuse features with different granularity more efficiently and explore the backbone with different depths.

This experiment is launched on the premise of NWPU-VHR-10 and RSOD data sets to test the TRCNet performance, manifesting the validity of our method. The second part is themed at related work, introducing the evolution of a target detection network in satellite remote sensing images and transformer structure development. Then, we discuss the proposed method, the general network structure, and principles related to the backbone network, neck, and detector in more detail. The experimental part is introduced later to elaborate on the results and analysis of the ablation experiment before the last summary of the full text.

2. Related Work

2.1 Evolution of Remote Sensing Target Detection

Up against the progress of CNN network architecture, the target detection algorithm performance has been greatly optimized. Various algorithms based on CNN network architecture have sprung up. Generally, given that whether the target detection algorithm has RPN or not, it is subsumed under single-stage target detection algorithms such as yolo series [34], SSD [35], Retinanet [12], etc., as well as two-stage ones such as RCNN [36], fast RCNN [37], faster RCNN [38], etc. Optical remote sensing image has the trait of scale diversity, visual angle particularity, high complexity of Beijing, smaller target than the background, and so on. However, the general target detection algorithms mentioned above are not specially designed for the problems of ORSI. Many workers have been working hard to solve these problems. The RP-Faster R-CNN framework [9] specially serves small target detection. Meanwhile, for the sake of importing detection compliance, deformable conversion layers [10] and R-FCN are united [11]. In this paper, the well-known Retinanet will be further improved to achieve better performance in remote sensing target detection tasks.

2.2 Transformer Structure

Transformer [13] was originally designed to solve NLP problems, with its unique self-attention mechanism used to model sequence input for long range, achieving great success in the NLP field. In recent years, researchers have spared no efforts to apply transformer modules to computer vision, which has proved that it also has the great potential [14-16] to rival or even surpass CNN in some fields. VIT was the first to use a transformer as the backbone network. For the sake of adapting to computer vision tasks, the input image is inclined to be uniformly divided into non-overlapping image blocks. Then, the transformer uses its multi-head attention mechanism to model the input image blocks in a long range and generates the feature map needed by downstream tasks. Although VIT [17] makes somewhat difference, it is fragile when encountering the multi-scale target detection and high-resolution tasks due to its inability to provide multi-scale feature maps and the high computational cost of a multi-head attention mechanism. PVT [18] effectively resolved these problems. It is a pyramid-structured transformer backbone with a spatial-reduction attention mechanism, which makes it still perform well confronted with multi-scale target detection and high-resolution tasks.

3. Proposed Method

3.1 Architecture Overview

The TRCNet architecture based on Retinanet is shown in Figure 1 which is subsumed under three main modules, that is, backbone, neck, and head. We input satellite remote sensing images into TRCNet and feature extraction will be carried out in the backbone part based on the transformer to obtain multi-scale feature maps C2, C3, C4, and C5. Then, C3, C4, and C5 will be input into Neck for a more detailed feature fusion operation. The neck is mainly composed of FPN with a carafe operator. After receiving the multi-scale feature map from the backbone, the FPN carafe module will carry out in-depth feature extraction plus detailed high-level and low-level feature fusion. Finally, P3, P4, P5, P6, and P7 as detailed feature maps will be generated. The Head acquire the input detailed feature map and then classify objects plus regress the bounding box. In this way, the final target detection results are output. The sections of backbone network and the Neck are explained in detail below.

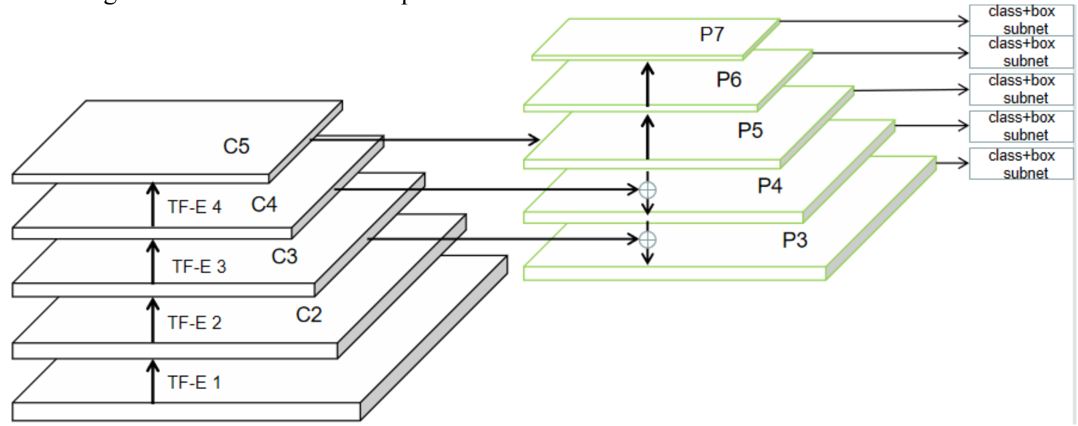

Figure 1. The Structure of TRCNet

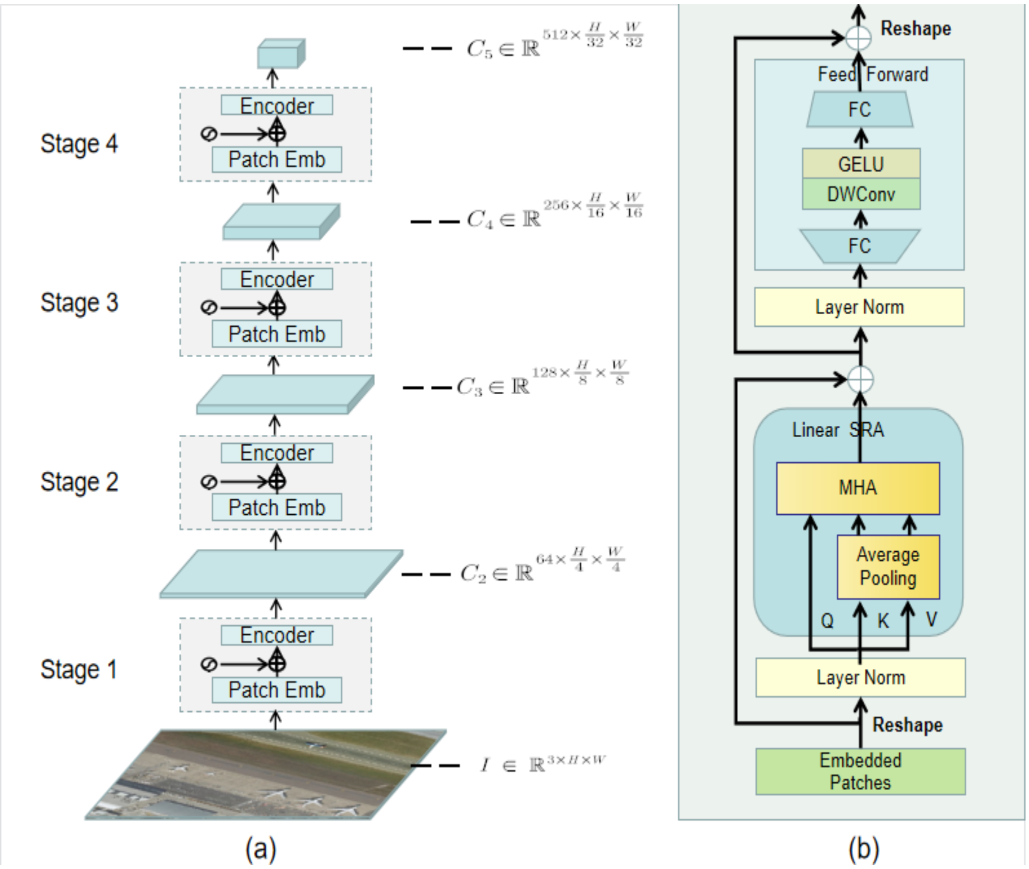
Stage 4
Stage 3
Stage 2
Stage 1
Encoder
Patch Emb
$C_5 \in \mathbb{R}^{512\times\frac{H}{32}\times\frac{W}{32}}$
$C_4 \in \mathbb{R}^{256\times\frac{H}{16}\times\frac{W}{16}}$
$C_3 \in \mathbb{R}^{128\times\frac{H}{8}\times\frac{W}{8}}$
$C_2 \in \mathbb{R}^{64\times\frac{H}{4}\times\frac{W}{4}}$
$I \in \mathbb{R}^{3\times H\times W}$
(a)
Reshape
Feed Forward
FC
GELU
DWConv
FC
Layer Norm
Linear SRA
MHA
Average Pooling
Q
K
V
Layer Norm
Reshape
Embedded Patches
(b)

Figure 2. Framework of PVTv2 Backbone. (b): Framework of PVTv2 Blocks

3.2 Backbone

Since the objects in ORSI are chaotic with obviously various size, the contrast between the background and objects is small, etc., and the interference to objects is serious. Therefore, how to detect objects accurately in ORSI is worthy of thinking [39]. Traditional CNN has a limited receptive field, so the global information acquisition in the target recognition task of ORSI takes time. Although the receptive field of CNN can be expanded through stacking depth and pooling operation, it will lead to the degradation of small target detection performance. To solve these problems, we introduce a transformer-based backbone-PVTv2 [19], as shown in Figure 2 (a). The transformer can adaptively extract local/global context information and model flexibly [20-22].

Similar to Retinanet's traditional Resnet 50-based hierarchical backbone, our transformer-based PVTv2 backbone consists of four stages outputing multi-scale feature maps. There is a Conv1 module before the first phase. First of all, the input image is preprocessed and the input image as $I \in \mathbb{R}^{3\times H\times W}$.

All stages adopt a similar structure of PVTv2blocks. We used PVTv2blocks with depths of 3, 4, 6, and 3 in the first, second, third, and fourth stages, as shown in Figure 2 (b). Different from the traditional transformer, each PVTv2blocks consists of a linear spatial reduction attention layer (LSRA), an overlapping patch embedding layer, and a conventional feed-forward layer (CFFL). $C_2 \in \mathbb{R}^{64\times\frac{H}{4}\times\frac{W}{4}}$ is output in the first stage, $C_3 \in \mathbb{R}^{128\times\frac{H}{8}\times\frac{W}{8}}$ the second, $C_4 \in \mathbb{R}^{256\times\frac{H}{16}\times\frac{W}{16}}$ the third, and $C_5 \in \mathbb{R}^{512\times\frac{H}{32}\times\frac{W}{32}}$ the fourth.

$$C_2 = \mathcal{F}^{3}_{pvt}(I)$$

$$C_3 = \mathcal{F}^{4}_{pvt}(C_2)$$

$$C_4 = \mathcal{F}^{6}_{pvt}(C_3)$$

$$C_5 = \mathcal{F}^{3}_{pvt}(C_4)$$

$$e.g., \mathcal{F}^{1}_{pvt} = \mathcal{F}_{CFFL}(\mathcal{F}_{LRSA}(\mathcal{F}_{linear}(X)) \oplus X)$$

Where $\mathcal{F}^{i}_{pvt}(\cdot)$ indicates a series of operations of i PVTv2blocks, $\mathcal{F}_{linear}(\cdot)$ linear project operation, $\mathcal{F}_{LRSA}(\cdot)$ the function of LSRA, and $\mathcal{F}_{CFFL}(\cdot)$ the operation of CFFL. With similarities to MHA, our SRA receives Q (a query), K (a key), and V (a value), with $Q,K,V \in \mathbb{R}^{(HW)\times C}$ respectively.

$$\mathcal{F}_{LSRA}(Q,K,V) = \mathcal{F}_{cat}(head_0,...head_N)W^A$$

$$head_j = \mathcal{F}_{atten}(QW_j^Q, \mathcal{F}_{avg}(K)W_j^K, \mathcal{F}_{avg}(V)W_j^V)$$

Where the spatial height and width of the input before the attention operation are represented by H and W. At the same time, C means the patch embedding dimension, and N the head number of self-attention detection, W^A, W_j^Q, W_j^K, W_j^V the weights of linear projection operation. The spatial average pooling operation can be demonstrated by $\mathcal{F}_{avg}(\cdot)$, while $\mathcal{F}_{cat}(\cdot)$ signifies the feature concatenation [59]. $\mathcal{F}_{atten}(\cdot)$ acts as the following self-attention function:

$$\mathcal{F}_{atten}(q,k,v) = soft\max\left(\frac{qk^T}{\sqrt{d}}\right)v$$

Where the channel number of various detection heads are signified by d = C/N. Average pooling is taken advantaged by LSRA to reduce the size of the scale. The application of LSRA can reduce the computational/memory burden compared to MHA. In this way, the transformer block is qualified to extract long-distance dependencies with global receptive fields in essence. Apart from that, a 3 × 3 depth-wise convolution is incorporated with GELU [60] activation layer by the PVTv2 block into CFFL between two entirely connected layers, in which CFFL is denoted as

$$\mathcal{F}_{CFFL}(X) = \mathcal{F}_{FC}(\mathcal{F}_{GELU}(\mathcal{F}_{DW}(\mathcal{F}_{FC}(X)))) \oplus X$$

Where the operation of a entirely connected layer is denoted by $\mathcal{F}_{FC}(\cdot)$, and the 3 × 3 depth-wise convolution $\mathcal{F}_{DW}(\cdot)$.

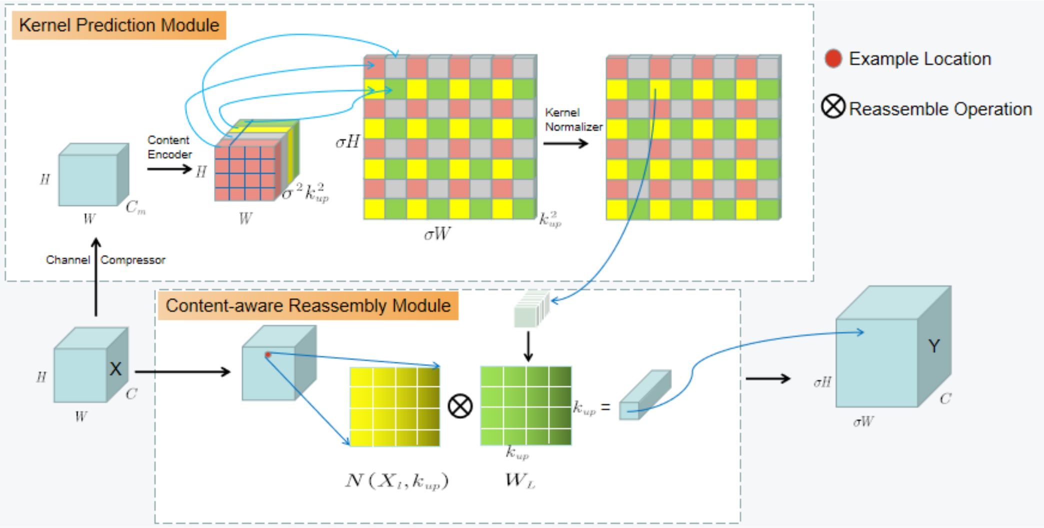

Figure 3. The Overall Framework of CARAFE Composed by Kernel Prediction Module and Content-Aware Assembly Module

3.2.1 Neck

Feature up-sampling is vital for multi-scale icon detection. After proposing the feature pyramid, it is more and more common to sample high-level features and fuse them with low-level ones. However, the traditional up-sampling method fails to use the feature map semantics, which limits the feature fusion potential. Decomposition obtains the up-sampling kernel through the network. Although it uses semantics, it introduces a lot of parameters and calculations and applies the same kernel at every position of the feature, so it fails to use the semantics of the feature graph efficiently. However, the CARAFE operator possesses a large receptive field and the up-sampling kernel is pertinent to the feature map semantics, which strengthens the effect of multi-scale target detection after fusing multi-level features without introducing too many parameters and calculations [8].

Therefore, we retain the Neck of the traditional Retinanet and adopt the FPN structure to fuse the multi-level feature map after sampling. We introduce the Carafe operator in the up-sampling to improve their fusion effect.

Carafe includes two steps as shown in figure 3. Step 1 is to generate an up-sampling kernel premised on the input feature map's semantics. Step 2 is to check the features of the input feature map taking into account the generated up-sampling to carry out up-sampling reorganization.

Suppose we input a feature map $X \in \mathbb{R}^{C\times H\times W}$ and the up-sampling ratio is $\sigma \in \mathbb{Z}^{+}$. A new feature map $Y \in \mathbb{R}^{C\times \sigma H\times \sigma W}$ will be generated after passing the Carafe operator. It is expressed as follows by the mathematical formula:

$$W_L = \psi(N(X_l, k_{encoder}))$$

$$Y_L = \phi(N(X_l, k_{up}), W_L)$$

Where L and l are target positioning, $L=(i^L, j^L)$ of the Y, $l=(i^l, j^l)$ is the mapped source location on X, and $i^l = \lfloor i^L/\sigma \rfloor, j^l = \lfloor j^L/\sigma \rfloor$. The $N(X_l, k)$ indicates the neighborhood of size $k \times k$ centered on X_l. As for ϕ, ψ, we will show the details later.

3.2.2 Kernal prediction module

The kernel prediction module can generate an up-sampling kernel based on the semantics of the input feature map. Every source position on X can correspond to the σ^2 target position on Y. Each target location has an up-sampling core of size $k_{up} \times k_{up}$.

In order to generate the up-sampling kernel, the first step is to mitigate the input feature channel from C to C_m via the 1×1 convolution, fastening the operation after reducing the channel. Then, the convolution kernel with kernel size $k_{encoder} \times k_{encoder}$ and channels = $\sigma^2 k_{up}^2$ is used for convolution operation to generate up-sampling kernels. Each $k_{up} \times k_{up}$ up-sampling kernel is normalized in space by softmax.

C-B: Content-Aware Restructuring Up-Sampling Kernel

For each up-sampling kernel W_L, the content-aware restructuring module will recombine through ϕ and calculate with the input feature map to get the up-sampling one, which is generally a weighting operator. With regard to a L (target location) and $N(X_l, k_{up})$ (the corresponding square region) centered at $l=(i,j)$, the following mathematical formula demonstrates the calculation:

$$Y_L = \sum_{n=-r}^{r}\sum_{m=-r}^{r} W_{L(n,m)}\cdot X_{(i+n,j+m)}$$

Where $r=\lfloor k_{up}/2 \rfloor$

3.2.3 Neck Finishing Process

Firstly, three feature maps C5, C4, and C3 from the backbone are received. Moreover, the three feature maps are up-sampled, feature extracted, and feature fused to obtain P7, P6, P5, P4, and P3 for downstream target detection, which is expressed by the mathematical formula:

$$D_5 = \mathcal{F}^1_{Conv2\_1}(C_5)$$

$$D_4 = \mathcal{F}^1_{Conv2\_2}(C_4)$$

$$D_3 = \mathcal{F}^1_{Conv2\_3}(C3)$$

Where $\mathcal{F}^1_{Conv2}(\cdot)$ refers to the Conv2 module, containing a conversion layer (kernel size = 1 × 1, stride size = 1, channels = 256).

$$P_5 = \mathcal{F}^1_{Conv5}(D_5)$$

$$P_6 = \mathcal{F}^1_{Conv6}(P_5)$$

$$P_7 = \mathcal{F}^1_{Conv7}(P_6)$$

Where $\mathcal{F}^1_{Conv5}(\cdot)$ refers to Conv5 module containing a conversion layer (kernel size = 3 × 3, stride size = 1, channels = 256). $\mathcal{F}^1_{Conv6}(\cdot)$ refers to Conv6 module containing a conversion layer (kernel size = 3 × 3, stride size = 2, channels = 256). $\mathcal{F}^1_{Conv7}(\cdot)$ refers to Conv7 module containing a conversion layer (kernel size = 3 × 3, stride size = 2, channels = 256).

$$P_4 = \mathcal{F}^1_{Conv4}(\mathcal{F}^1_{Carafe}(D_5)\oplus D_4)$$

$$P_3 = \mathcal{F}^1_{Conv3}(\mathcal{F}^1_{Carafe}(D_4)\oplus D_3)$$

Where $\mathcal{F}^1_{Conv4}(\cdot)$ refers to Conv4 module containing a conversion layer (kernel size = 3 × 3, stride size = 1, channels = 256). $\mathcal{F}^1_{Conv3}(\cdot)$ refers to Conv3 module containing a conversion layer (kernel size = 3 × 3, stride size = 2, channels = 256). $\mathcal{F}^1_{Carafe}(\cdot)$ refers to Carafe up-sampling.

3.2.4 Head and Loss

The target detection probe and the loss we use will be introduced here. Same as the traditional Retinanet, the target detection probe used in this paper is a weight-sharing predictor based on convolution operation. It is divided into two branches, which respectively predict each anchor's category and the regression parameters of the target bounding box.

Positive and negative samples is the same as Retinanet in matching strategy usage. Comparing each anchor with the pre-labeled GT box, the positive sample IOU is more than 0.5. If the IOU value of the anchor and all GT boxes is less than 0.4, it is negative. The rest are discarded.

Total loss consists of classification loss and regression loss. Both positive and negative samples will calculate the classification loss. But only the positive ones will be calculated for the regression loss.

$$Loss = \frac{1}{N_{POS}}\sum_i L^i_{cls} + \frac{1}{N_{POS}}\sum_j L^j_{reg}$$

Where L_{cls} indicates the Sigmoid Focal loss. L_{reg} indicates the L1 loss. N_{POS} denotes the number of positive samples. i all the positive samples. j all the negative samples.

4. Experiment

4.1 Dataset

The NWPU VHR-10 data set published by Northwestern University in 2014 [23, 24, 25] contains 10 categories of objects, that is, aircraft, ships, storage tanks, baseball diamonds, tennis courts, basketball courts, ground orbital fields, ports, bridges, and vehicles. The dataset contains 800 very high resolution (VHR) RSI derived from the Google Earth and Vaihingen datasets, which are annotated by experts in person. We randomly divided 640 pictures into training sets and the remaining 160 pictures were test sets.

RSOD is an open target detection dataset used for target detection in RSI. There are four kinds of objects, including airplanes, fuel tanks, sports fields, and overpasses. It includes 4,993 aircraft in 446 images, 191 playgrounds in 189 images, 180 overpasses in 176 images, and 1,586 fuel tanks in 165 images. 885 pictures have been divided as training sets and the remaining 216 pictures test sets.

4.2 Implementation Details

We train the TRCNet by using PyTorch on a PC with 4 kernels Intel (R) Xeon (R) Silver 4110 CPU @ 2.10 GHz, 16-GB RAM, and an NVIDIA GTX 2080Ti GPU. We adopt data argumentation - Random flip before training. In the training, this network shoulders the pretrained weights of the backbone and the remaining parameters are randomly initiated by Xavier. Besides, the mesoscale is (1,000, 600) and the keep ratio is true, the max epoch is 72, the batch size is 4, and the optimizer is Adam W. Meanwhile, the learning rate and the weight decay are both 10^{-4}. We adopt the operation of warm-up. VOC2007 11-point metric [33] is applied to evaluate the proposed method's performance. s show in Figure 4.5.6.

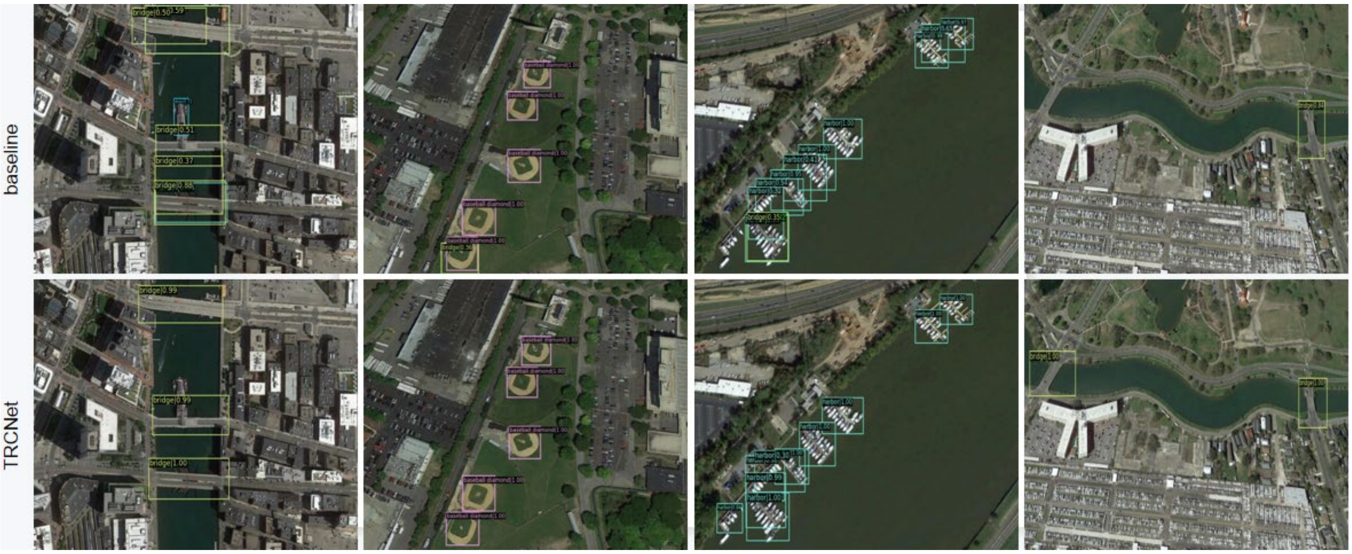

Figure 4. Detection Examples of the Proposed Method and Baseline in NWPU VHR-10

1

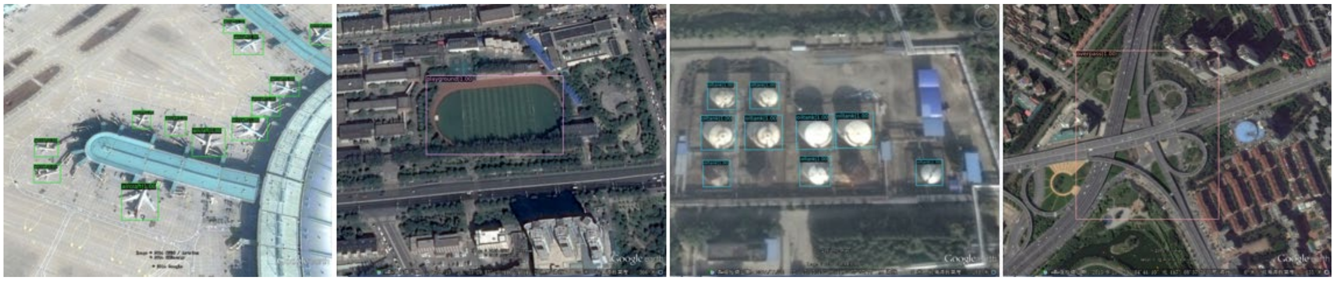

Figure 5. Detection Examples of the Proposed Method in RSOD**2**

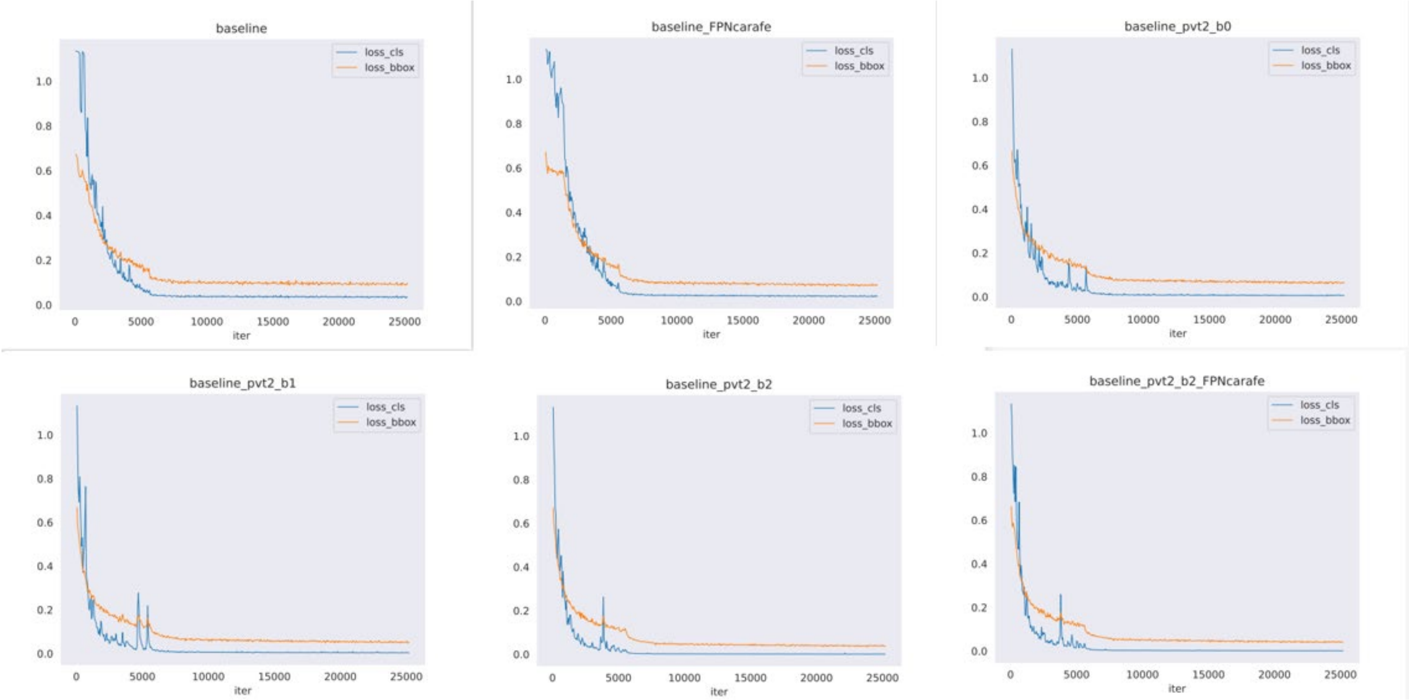

Figure 6. The Loss Curve of Training Process

4.3 Comparison Results with the Latest Methods

Its Performance will be evaluated quantitatively in this body. We compared it with some advanced methods on the NWPU VHR-10 data set, pertinent results are in Table 1. We use several general target detection algorithms (FCOS [26], R-FCN [27], Cascade R-CNN [28], AugFPN [29]), and methods of remote sensing image object detection (MS-FF [30], HRBM [31], SHDET [32]. The experienced results (mAP, AP) in the below tables are converted to percent (%).

Table 1 Accuracy Comparison of Different Object Detection Methods on The NWPU VHR-10 Dataset1

Method	Map	AP_1	AP_2	AP_3	AP_4	AP_5	AP_6	AP_7	AP_8	AP_9	AP_10
R-FCN	87.74	99.80	80.82	90.48	97.88	90.69	72.38	98.99	87.18	70.44	88.62
Cascade R-CN	89.15	99.90	80.05	90.67	98.06	89.49	76.71	98.21	80.76	87.67	**89.92**
AugFPN	89.16	90.91	71.83	90.62	98.59	90.77	84.24	99.24	91.66	84.71	89.06
FCOS	85.84	90.47	73.72	90.36	98.94	89.38	80.82	96.74	87.91	61.92	88.16
MS-FF	85.64	95.79	72.50	70.90	97.83	85.62	97.20	98.82	92.40	81.74	64.64
HRBM	87.12	99.70	**90.80**	90.61	92.91	90.29	80.13	90.81	80.29	68.53	87.17
SHDet	90.04	**100.00**	81.36	90.90	98.66	90.84	82.57	98.68	91.11	76.43	89.82
XXXNet	**92.70**	90.90	77.50	**96.60**	90.70	90.20	**99.00**	**99.80**	**98.40**	**95.10**	88.90

AP1 to AP10 Successfully Correspond to Airplanes, Ships, Storage Tanks, Baseball Diamonds, Tennis Courts, Basketball Courts, Ground Track Fields, Harbor, Bridges, and Vehicles

Table 2. Ablation Study of Pvtv2 Backbone2

Method	mAP
Baseline	84.3%
Baseline + PVTv2_b0	89.5%

Table 3. Ablation study of The Depth of Pvtv2 Backbone 3

Method	mAP
Baseline	84.3%
Baseline + PVTv2_b0	89.5%
Baseline + PVTv2_b1	90.0%
Baseline + PVTv2_b2	91.7%

Table 4. Ablation Study of the Carafe4

Method	mAP
Baseline	84.3%
Baseline + Carafe	87.0%

Table 5. Ablation Study of the Carafe and PVTv2 Backbone5

Dataset	Method	mAP
NWPU-VAR-10	Baseline	84.3%
	Baseline + PVTv2_b2 + Carafe	92.7%
RSOD	Baseline	91.3%
	Baseline + PVTv2_b2 + Carafe	92.9%

4.4 Ablation Study

To test whether the TRCNet works or not, we designed four ablation experiments premised on the NWPU-VHR-10 data set, and finally tested the TRCNet validity premised on the RSOD data set. We use Rtinanet-Resnet 50 as the baseline.

4.4.1 Analysis of TRCNet Backbone

We replaced the traditional Retinanet backbone based on Resnet50 with PVTv2-b0 based on the transformer. After training epoch=72 on the NWPU-VHR-10 data set, we can find in Table 2 that the mAP of baseline is 0.843, while the mAP of the baseline_PVTv2_b0 is 0.917, which increases by 7.4%. This is because the transformer has a large receptive field, which can be used for long-range modeling. It is more efficient than cnn structure to extract features in satellite remote sensing images when the background interference is large and the distribution of objects is messy.

4.4.2 Analysis of Different Depths of TRCNet Backbone

We replaced Retinanet's backbone with PVTv2 modules of different depths to explore the performance of PVTv2 modules of different depths. We use PVTv2_b0, PVTv2_b1, and PVTv2_b2 respectively. The mAP of PVTv2_b0, PVTv2_b1, and PVTv2_b2 in Table 3 are 0.895, 0.900, and 0.917 respectively. In the wake of increasingly advanced network depth, the mAP heralds an upward trend. Apart from that, the backbone feature extraction is enhanced and the obtained feature information is more abundant. Considering the size of network parameters, this paper only goes deep into b2.

4.4.3 Analysis of FPN_carafe

On the basis of the baseline, we introduce the Carafe operator into the up-sampling feature fusion module of FPN, train it, and compare it with the baseline. Through Table 4, we can find that after introducing the Carafe operator, the mAP of baseline _FPNcarafe is 0.870, which is 2.7% higher than baseline_PVTv2_b2. This is because the Carafe operator can guide feature fusion more efficiently when FPN is fused with up-sampling features. At the same time, the fused features obtained are more accurate and richer than those obtained by FPN fusion alone, which makes it possible to obtain better performance in multi-scale target detection.

4.4.4 Analysis of backbone and FPN_carafe

On the basis of the baseline, we introduce the Carafe operator into the up-sampling feature fusion module of FPN and replace its backbone with PVTv2_b2, training and comparing with baseline. Through Table 5, we can find that after introducing the Carafe operator and replacing backbone, the mAP of the baseline_PVTv2_b2_FPNcarafe on the NWPU-VAR-10 data set is 0.927, which is 8.4% higher than baseline. Meanwhile, that of baseline_PVTv2_b2_FPNcarafe on the RSOD data set is 0.929, which is 1.6% higher than the baseline. It shows that the introduction of PVTv2_b2 and carafe can play a role at the same time.

4.4.5 Figures that still need to be supplemented

(1) The comparison figure between the NWPU baseline model and XXNET detection results is generally divided into two rows. The previous row is four figures of baseline, and the next row is the corresponding detection figure with a better XXNET effect than the baseline.

(2) It's enough to give 4 pictures with a score of about 0.9 for the ROSD test result diagram, and the box test is relatively accurate.

(3) The loss curve graph

Section that needs to be added:

Compared with the state-of-the-art

Comparison with other networks.

5. Conclusion

This paper explores the feasibility of a target detection algorithm based on a new transformer structure in RSI with serious background interference, different target sizes, and uneven distribution of geospatial objects. Targeting at further boosting the accuracy of remote sensing target detection, the weakening performance caused by insufficient global modeling ability of traditional CNN-based model and the low efficiency of feature fusion caused by uniform up-sampling of the FPN network is solved. In the proposed TRCNet, we introduce a hierarchical PVTv2_b2 module based on a transformer as a backbone to extract features, so as to obtain more accurate and richer feature maps than the backbone based on CNN. Then Carafe operator is introduced in the multi-level features fusion of the FPN network. This operator can use the semantics in the upper feature map to guide the up-sampling process, instead of uniform up-sampling, which makes the feature fusion of the FPN network efficient and accurate, perfecting the performance of multi-scale target detection. Finally, compared with Retinanet, the mAP of TRCNet on NWPU-VHR-10 and RSOD increased by 8.4% and 1.7% respectively. Furthermore, we also compare the results of the network on NWPU VHR-10 with other advanced networks, which proves its advanced nature.

References

[1] J. Niemeyer, F. Rottensteiner, and U. Soergel, “Contextual classification of lidar data and building object detection in urban areas,” ISPRS J. Photogrammetry Remote Sens., vol. 87, pp. 152–165, 2014.

[2] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, “Building detection in very high-resolution multispectral data with deep learning features,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., pp. 1873–1876, 2015.

[3] Z. Chen, T. Zhang, and C. Ouyang, “End-to-end airplane detection using transfer learning in remote sensing images,” Remote Sens., vol. 10, Art. no. 139., 2018.

[4] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. Conf. Adv. Neural Inform. Process. Syst., pp. 379–387, 2016.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 779–788, 2016.

[6] Q. Zhang et al., “Dense attention flfluid network for salient object detection in optical remote sensing images,” IEEE Trans. Image Process., vol. 30, pp. 1305–1317, 2021.

[7] X. Zhou et al., “Edge-Guided Recurrent Positioning Network for Salient Object Detection in Optical Remote Sensing Images,” in IEEE. Transactions on Cybernetics, doi: 10.1109/TCYB.2022.3163152.

[8] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy and D. Lin, “CARAFE: Content-Aware Reassembly of Features,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3007-3016, 2019, doi: 10.1109/ICCV.2019.00310.

[9] Xiaobing Han, Yanfei Zhong, and Liangpei Zhang. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. Remote Sensing, 9(7), 666, 2017.

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. CoRR, abs/1703.06211, 1(2), 3, 2017.

[11] Zhaozhuo Xu, Xin Xu, Lei Wang, Rui Yang, and Fangling Pu. Deformable convnet with aspect ratio constrained nms for object detection in remote sensing imagery. Remote Sensing, 9(12), 1312, 2017.

[12] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, “Focal Loss for Dense Object Detection,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318-327, 1 Feb. 2020. doi: 10.1109/TPAMI.2018.2858826.

[13] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., pp. 5998–6008, 2017.

[14] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted Windows,” 2021, arXiv:2103.14030.

[15] W. Wang, L. Yao, L. Chen, D. Cai, X. He, and W. Liu, “Crossformer: A versatile vision transformer based on cross-scale attention,” 2021, arXiv:2108.00154.

[16] X. Dong et al., “CSWin transformer: A general vision transformer backbone with cross-shaped Windows,” 2021, arXiv:2107.00652.

[17] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929

[18] W. Wang et al., “Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 548-558, doi: 10.1109/ICCV48922.2021.00061.

[19] “PVT v2: Improved Baselines with Pyramid Vision Transformer,” arXiv:2106.13797

[20] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), pp. 1–15, 2021.

[21] W. Wang et al., “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 568–578, Oct. 2021.

[22] W. Wang et al., “PVT v2: Improved baselines with pyramid vision transformer,” Comput. Vis. Media, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[23] Gong Cheng, Junwei Han, Peicheng Zhou, Lei Guo. Multi-class geospatial object detection and geographic image classification based on collection of part detectors.

ISPRS Journal of Photogrammetry and Remote Sensing, 98: 119-132, 2014.

[24] Gong Cheng, Junwei Han. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 117: 11-28, 2016.

[25] Gong Cheng, Peicheng Zhou, Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 54(12), 7405-7415, 2016.

[26] Tian Z, Shen C, Chen H, Fcos TH. Fully convolutional one stage object detection. In: Proceedings of the IEEE international conference on computer vision, pp 9627–9636, 2019.

[27] Dai J, Yi L, He K, Sun J. R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp 379–387, 2016.

[28] Cai Z, Vasconcelos N. Cascade r-cnn: High quality object detection and instance segmentation, IEEE Trans Pattern Anal Mach Intell 1–1, 2019.

[29] Guo C, Fan B, Zhang Q, Xiang S, Pan C. Augfpn: Improving multi-scale feature learning for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12595–12604, 2020.

[30] Zhang W, Jiao L, Liu X, Liu J. Multi-scale feature fusion network for object detection in vhr optical remote sensing images. In: IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium, pp 330–333. IEEE, 2019.

[31] Li K, Cheng G, Bu S, You X (2018) Rotation-insensitive and context-augmented object detection in remote sensing images. IEEE Trans Geosci Remote Sens 56(4), 2337–2348, 2018.

[32] Zhu, D., Xia, S., Zhao, J. et al. Spatial hierarchy perception and hard samples metric learning for high-resolution remote sensing image object detection. Appl Intell 52, 3193–3208, 2022. https://doi.org/10.1007/s10489-021-02335-0

[33] Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. Int JComput Vis 88(2), 303–338, 2010.

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unifified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 779–788, Jun. 2016.

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

[36] R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014. doi: 10.1109/CVPR.2014.81.

[37] R. Girshick, “Fast R-CNN,” 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[38] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[39] X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang and C. Yan, “Edge-Aware Multiscale Feature Integration Network for Salient Object Detection in Optical Remote Sensing Images,” in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-15, 2022, Art no. 5605315, doi: 10.1109/TGRS.2021.3091312.