Refining FFT-based Heatmap for the Detection of Cluster Distributed Targets in Satellite Images

Huan Zhang[1,3]
zhanghuan19@mails.tsinghua.edu.cn
Zhiyi Xu[2,3]
xuzhiyi@tsinghua.edu.cn
Xiaolin Han[1,3]
hxl15@tsinghua.org.cn
Weidong Sun[1,3]
wdsun@tsinghua.edu.cn

[1] Department of Electronic Engineering
Tsinghua University
Beijing, China

[2] China University of Geosciences (Beijing)
Beijing, China

[3] Institute for Ocean Engineering
Tsinghua University
Beijing, China

Abstract

The detection of cluster distributed targets in remotely sensed satellite images is a challenging task, as cluster is a common behavior of targets and adhesions between dense distributed targets often exist, which affect the accuracy of object detection seriously. However, the distinct distribution pattern of such cluster distributed targets in frequency domain has never been studied. In this paper, a refinement of FFT-based heatmap with multi-branches network for the detection of cluster distributed targets in the satellite images (termed as HeatNet) is proposed. More specifically, a refining method of the FFT-based heatmaps for different features in frequency domain and an attention-based feature extractor in frequency channel are proposed, to focus the attention and refine the salient regions for the cluster distributed targets. Additionally, as one complete system, a keypoint-based detection is adopted as the basic workflow to tackle with the adhesion, a scale-aware center area is conducted to tackle with the variation of scale, and an orientation discrimination is also utilized to eliminate the specificity of different targets. The effectiveness of our proposed method is validated on two public datasets, and the comparative experimental results with different state-of-the-arts object detection methods have demonstrated the superiority of this proposed method.

1 Introduction

The detection of cluster distributed targets in remotely sensed satellite images is a challenging task, because cluster is a common behavior of targets, and adhesions between dense distributed targets often exist in the satellite images. However, the distinct distribution pattern of such cluster distributed targets in frequency domain has never been studied deeply.

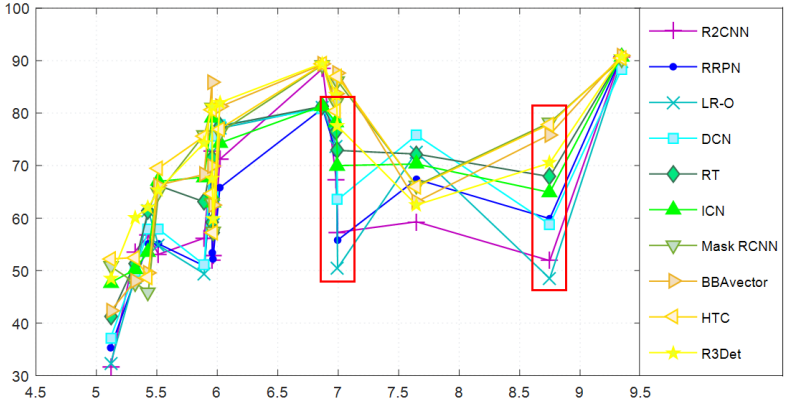


Figure 1: Relation between the combination index Ψ_i and the accuracy of object detection.

To show the above problem more clearly, 10 most representative detection methods from the Object Detection in Aerial Images (ODAI) challenge [1] have been analyzed, to estimate the influence of cluster distribution to the accuracy of object detection. Fig.1 depicts the relationship between the common impact factors of remote sensing object detection (including target size ρ, target number μ and aspect ratio ν) and the accuracy of object detection, in which each method is denoted by a different colored mark respectively. It is acknowledged that, for the learning based object detection, the detection accuracy increase with the increase of the combination index Ψ_i of these common impact factors, which can be formulated as Eq.(1), in accordance with the x-axis of Fig.1.

$$\Psi_i = log(\rho_i) \times (\mu_i)^{0.5} / f(\nu_i), i = 1, \ldots, N \tag{1}$$

$$f(\nu_i) = \begin{cases} |\nu_i - 1/N \times \Sigma\nu_i|, & |\nu_i - 1/N \times \Sigma\nu_i| \leq |\nu_i - 1/(1/N \times \Sigma\nu_i)| \\ (\nu_i - 1/(1/N \times \Sigma\nu_i))^2, & otherwise \end{cases} \tag{2}$$

However, the last two columns marked with red boxes indicate this argument is not always true, which means that there is other impact factor influences the accuracy of object detection, except for the existing impact factors.

The reason why the accuracy of object detection reduces at certain points, can be explained by the cluster distribution and the adhesion between dense distributed targets, which is particularly prominent for the common targets like ship and small-vehicle corresponding to the last two columns with red boxes in Fig.1. In the adhesion regions, borders of different targets overlapped with each other, and the salient regions being connected in the feature maps, which make it difficult to be focused and detected separately with the traditional object detection methods as shown in Fig.2. We can also see that, adhesion not only exists among the dense distributed targets, but also exists among dense distributed categories such as ship category and harbor category, and adhesion not only exists in the original images but also exists in the different feature maps, when a down-sampling operation is applied to the network convolution.

In addition, as pointed in [26], CNN models are more sensitive to low-frequency channel for general targets than high-frequency channel for dense distributed targets, and this can be analyzed from the frequency perspective which is very similar to the human visual system.

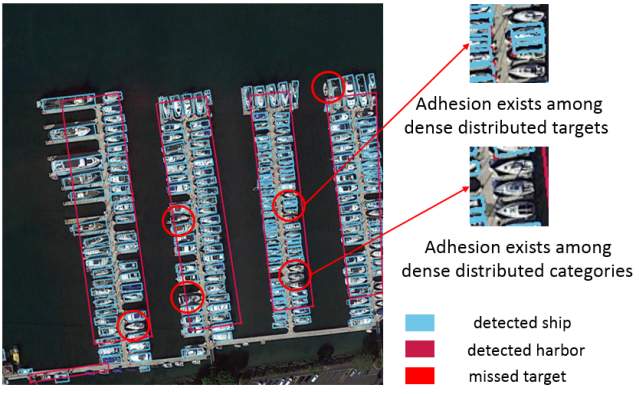


Figure 2: An example of adhesion phenomenon for the dense distributed ship targets in harbor. In which, blue boxes denote the detected ships, deep red boxes denote the detected harbors, and red circles denote the undetected ships.

According to the above discussions, a refinement of the FFT-based heatmap for the detection of cluster distributed targets in the satellite images (termed as HeatNet) is proposed in this paper, try to realize the detection of cluster distributed targets in the frequency domain. In detail, a refining of the FFT-based heatmaps for different features in frequency domain and an attention-based feature extractor in frequency channel is proposed, to focus the attention and refine the salient regions for the cluster distributed targets. Additionally, as one complete system, a keypoint-based detection is adopted as the basic workflow to tackle with the adhesion, a scale-aware center area is conducted to tackle with the variation of scale, and an orientation discrimination is also utilized to eliminate the specificity of different targets. The main novelties and contributions of our proposed HeatNet method are as follows:

(1) This is the first time that a new combination index of the common impact factors for the object detection has been formulated quantitatively, and the phenomenon of adhesion is explored explicitly according to the relationship between the combination index and the accuracy of object detection in satellite images.

(2) To our best knowledge, this is the first time that the distribution pattern of cluster distributed targets in the frequency domain has been studied for the object detection, and the FFT is introduced to refine the heatmaps rather than to accelerate CNN calculation.

(3) The framework of HeatNet consisting of scale-aware center area, orientation discrimination, refining FFT-based heatmaps in frequency domain and attention-based feature extractor in frequency channel, is an effective solution for the object detection in satellite images, especially for the dense distributed targets.

2 Related works

2.1 Object detection in satellite images

The methods for object detection in satellite images can be divided into two categories, anchor based detection and keypoint based detection. The early anchor based detection methods generally refer to the two-stage detection methods, which decompose the detection into region proposals generation and candidate boxes classification, such as R2CNN [14],

R2PN [36], ICN [1], FFA [9], RoI Transformer [6], Gliding Vertex [31], CenterMap [27] and ReDet [12] method. Then, anchor mechanism has also been extensively utilized in one-stage detection methods, to improve the accuracy of object detection, such as RetinaNet [18], SCRDet [33] and R3Det [32] method. But, due to the adhesion, object in the satellite images is easy to be missed by the non-maximum suppression (NMS) of anchors. In addition, the accuracy of anchor based methods depend on the anchor setting, which is inappropriate to tackle with the large aspect ratio targets.

The keypoint-based detection methods transform the object detection into predicting and grouping of the keypoints, such as CornerNet [15], CenterNet [37] and ExtremeNet [38] method. Although, the variant scale of targets has also been considered in BBAvector [34] and the arbitrary orientation of targets has been considered in O2-DNet [28], the cluster distribution of targets in the satellite images has never been studied in such kind of methods.

2.2 Fourier Transforms in deep learning

Fourier Transforms, especially the Fast Fourier Transform (FFT), has been firstly used in neural networks to identify the electrocardiogram signals of the heart [23] [10] [24]. In deep learning, FFT has been applied to speed up the computation of CNNs as in [8], [22] and [25]. In addition, FFT has also been applied to Recurrent Neural Networks (RNNs) to stabilize training, and to reduce the gradients exploding and vanishing problems [35]. Recently, FFT has been utilized in Transformer, to linearize the complexity of self-attention mechanism by leveraging random Fourier features [4], and the FNet [16] to replace the self-attention sublayer. But, the FFT hasn't been introduced for the task of object detection itself rather than the acceleration of CNNs.

2.3 Frequency domain learning

In recent years, frequency analysis has emerged in the deep learning. The frequency analysis has been introduced firstly into CNN for the JPEG transform [7]. In [30], DCT analysis has been utilized to identify and remove the trivial frequency components without decreasing the accuracy. In addition, frequency domain learning has also applied to model compression [3] and model pruning [19]. However, frequency domain has never emerged in the object detection of satellite images. Whereas, in this paper, we will focus on the distinct distribution pattern of cluster distributed targets in frequency domain, and take the keypoint based object detection method [34] as the basic workflow.

3 Method

The overall architecture of our proposed HeatNet method can be described as Fig.3. In which, the HeatNet method is consist of 3 parts, attention-based feature extractor, refinement of FFT-based heatmaps, orientation discrimination and bounding box determination. As the data flow, for a given input image with a dimension of $w \times h$, it will be sent to the attention-based feature extractor in frequency channel. Then, the outputs of feature extractor will be sent into the 4 branches in parallel, including the refinement of FFT-based heatmaps $B^H \in R^{H/s \times W/s \times C_H}$, the offset of center points $B^O \in R^{H/s \times W/s \times 2}$, the orientation discrimination $B^D \in R^{H/s \times W/s \times 1}$ and the bounding box determination $B^B \in R^{H/s \times W/s \times 10}$.

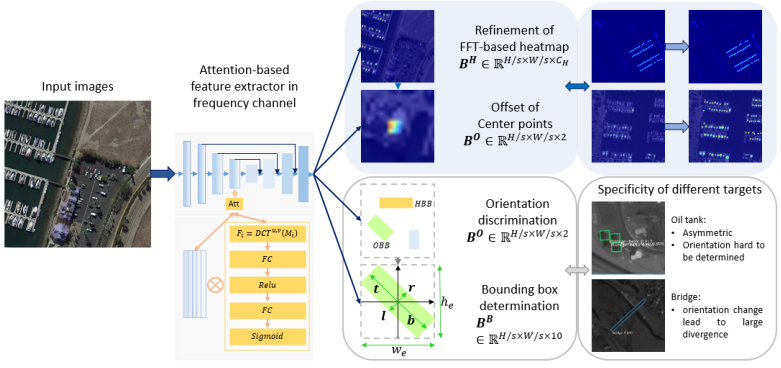


Figure 3: The architecture of our proposed HeatNet method, which is consist of 3 parts, attention-based feature extractor, refinement of FFT-based heatmaps, orientation discrimination and bounding box determination.

3.1 Attention-based feature extractor in frequency channel

As pointed in [30], CNN models are more sensitive to general low-frequency targets than high-frequency targets with dense distribution. To refine the feature extractor from the frequency perspective for the cluster distributed targets, an attention-based feature extractor in frequency channel is introduced as the following:

$$\Psi_{Att} = sigmoid(fc(F(M))) \tag{3}$$

In existing attention methods, especially in the channel attention methods, the general preprocessing of global average pooling (GAP) may lead to the information inadequacy problem. As GAP is the lowest frequency component of discrete cosine transforms (DCT) [26], in this method, more frequency components, $(F = \bigoplus([F_0, F_1, , F_{n-1}]), F \in R^C, \bigoplus$ denotes concate), are introduced into channel attention to solve the problem of information inadequacy. These frequency components are transformed by $F_i = DCT^{u,v}(M_i)$, where M_i is the split parts of feature map M along the channel dimension, i.e. $M = [M_0, M_1, \ldots, M_{n-1}]$, $M_i \in R^{H \times W \times c'}$, $i \in 0, 1, \ldots, n-1$, $c' = C/n$. Then, DCT can be formulated by the weighted sum as following:

$$F_i = DCT^{u,v}(M_i) = \Sigma_{h=0}^{H-1} \Sigma_{w=0}^{W-1} M_{:,h,w}^{i} \omega_{h,w}^{u,v} \tag{4}$$

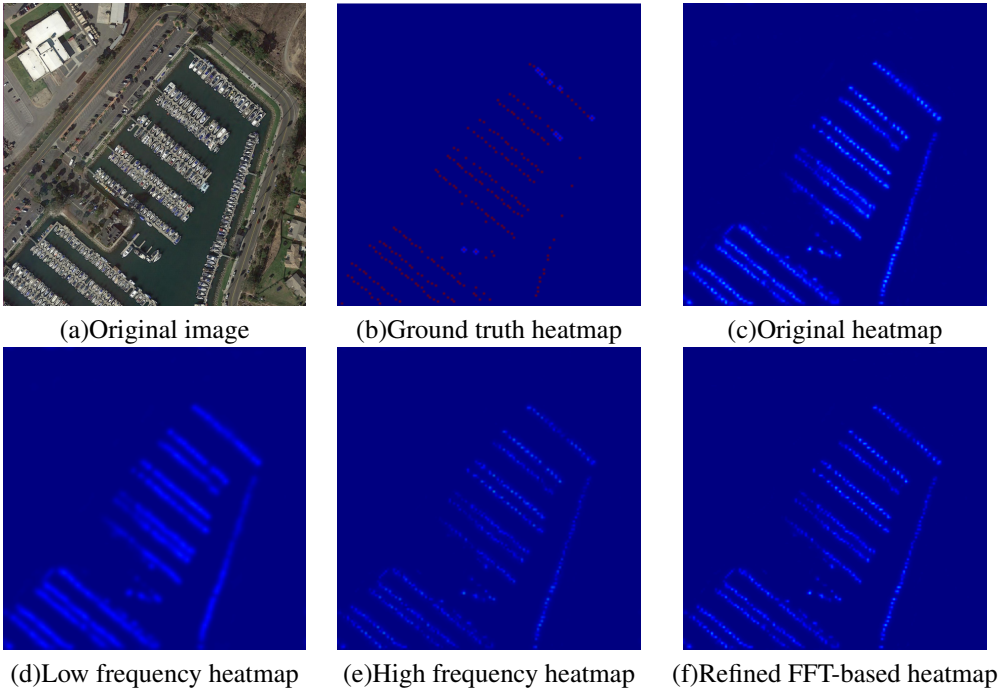
In which, the frequency component ω represents the 2D DCT:

$$\omega_{h,w}^{u,v} = cos(\pi h / H(u + 1/2)) cos(\pi w / W(v + 1/2)) \tag{5}$$

After acquiring the massive frequency components, the influence of each frequency component on attention can be acquired separately, the top k frequency components, $F = \bigoplus([F_0, F_1, \ldots, F_k]), F \in R^C$ will be selected and reserved as in [26]. The above attention-based feature extractor in frequency channel can focus and refine the feature maps from the frequency perspective for the cluster distributed targets.

3.2 Refinement of FFT-based heatmaps in frequency domain

The outputs of feature extractor will be sent into the 4 branches in parallel, among which the first branch predicts the heatmaps $B^H \in R^{H/s \times W/s \times C_H}$, here, s denotes the stride, and channel



(a)Original image (b)Ground truth heatmap (c)Original heatmap

(d)Low frequency heatmap (e)High frequency heatmap (f)Refined FFT-based heatmap

Figure 4: The refinement of FFT-based heatmaps.

C_H denotes the number of categories. The heatmap of each channel $B_i^H, i = 1,2,\ldots,C_H$, is passed through a sigmoid function, to acquire the predicted center points $c_i = (c_x, c_y), c_i \in [0,1]^{(H/s \times W/s)}$, which are taken as the targets confidence.

Due to the adhesion not only exists among dense distributed targets but also exists among dense distributed categories in the satellite images, in the adhesion regions, the borders of different targets may be overlapped to a certain degree, so the salient regions will be connected in the feature maps and heatmaps, which will be further aggravated in the following down sampling operations. To reduce the adhesion of the cluster distributed targets, here, a refinement of the FFT-based heatmaps in frequency domain is proposed. In detail, we deploy the 2D FFT on each heatmap to acquire the spectrum in frequency domain $S_i^H = FFT(B_i^H)$, $i = 1,2,\ldots,C_H$. Then, corresponding Gaussian filters with a kernel radius of 12 are conducted to divide the spectrum into the spectrum of low frequency $S_i^{H_l} = F(S_i^H)$ and the spectrum of high frequency $S_i^{H_h} = F(S_i^H)$ followed by the IFFT of both spectrums. As can be observed with Fig.4, for the cluster distributed targets, the heatmap of high frequency spectrum in Fig.4(e), i.e. $B_i^{H_h} = IFFT(S_i^{H_h})$, is less adhesive compared with the original heatmap, while the heatmap of low frequency spectrum in Fig.4(d), i.e. $B_i^{H_l} = IFFT(S_i^{H_l})$, is more adhesive. Therefore, the original B_i^H is replaced with a weighted $B_i^{H_h}, B_i^{H_l}$, as the final refined FFT-based heatmap shown in Fig.4(f), to separate the salient regions from each other more efficiently:

$$B_i^{'H} = \alpha_o B_i^H + \alpha_h B_i^{H_h} - \alpha_l B_i^{H_l}, i = 1,2,\ldots,C_H \tag{6}$$

In addition, the ground-truth heatmap $B_{GT}^H \in R^{H/s \times W/s \times C_H}$ is calculated with the Gaussian distribution. As claimed in [21], encoding more training samples from annotated boxes

is similar to increasing the batch size, which contributes to improving the detection accuracy and speeding up the convergence. And, the box size-adaptive standard deviation δ , is utilized to modify the center area as in [15] [37]. It worth mentioned that, the scale aware center area is friendly for the detection of cluster distributed small targets. Furthermore, to solve the imbalance between the limited positive center points and numerous negative points, the focal loss is utilized to train the heatmap as in [37].

During the center point prediction, the integer center points c down-scaled from the input images to the output heatmaps, will generate the floating center points $\hat{c}$. Therefore, the offset $B^O \in R^{H/s\times W/s\times 2}$ between c and $\hat{c}$ is predicted on the second branch to recover the discretization error and optimized by the smooth L1 loss.

3.3 Orientation discrimination and bounding box determination

There is specificity of different objects in the satellite images in shape and orientation, such as the oil tank and the bridge. The former is asymmetric, the orientation of which is hard to be determined. As for the latter, a fractional orientation change can lead to a large divergence on the detection accuracy, for the sensitivity of Intersection-over-Union (IoU) between the predicted bounding box and the ground-truth bounding box. To eliminate the specificity of different objects, here, an orientation discrimination is utilized, i.e. the bounding box (BB) of objects are firstly discriminated into horizontal bounding box (HBB) or oriented bounding box (OBB) as a divide-and-conquer strategy, which can be formulated as $B^D \in R^{H/s\times W/s\times 1}$. The ground-truth orientation B^D_{GT} is calculated by the IoU between the BB and the HBB, and the orientation discrimination is optimized by binary cross-entropy loss as in [34].

$$B^D_{GT} = \begin{cases} 0\ (HBBs), & IoU(BB,HBB) \geq 0.95 \\ 1\ (OBBs), & otherwise \end{cases} \tag{7}$$

After the discrimination of horizontal and oriented targets, the bounding box is utilized to describe the target regions, $B^B \in R^{H/s\times W/s\times 10}$, which includes the top points $t = (t_x, t_y)$, the bottom points $b = (b_x, b_y)$, the left points $l = (l_x, l_y)$, the right points $r = (r_x, r_y)$, and the parameters of external horizontal bounding box w_e and h_e [34]. The smooth L1 loss is utilized to regress the box parameters $B^B_i = [t, r, b, l; w_e, h_e]$ at the center points as in Fig.3.

The top-left (tl), top-right (tr), bottom-right (br) and bottom-left (bl) points of the bounding boxes (BB) are taken as decoded points from the $B^B_i = [t, r, b, l; w_e, h_e]$. In particular, for a center point c, the decoded HBB points can be acquired by:

$$tl = (c_x - w_e/2, c_y - h_e/2),\ tr = (c_x + w_e/2, c_y - h_e/2) \tag{8}$$

$$bl = (c_x - w_e/2, c_y + h_e/2),\ br = (c_x + w_e/2, c_y + h_e/2)$$

And the decoded OBB points can be acquired by:

$$tl = (t + l) + c,\ tr = (t + r) + c,\ bl = (b + l) + c,\ br = (b + r) + c \tag{9}$$

4 Experiments

In this section, the datasets used in the following experiments and the implement of our proposed HeatNet method will be introduced firstly. Then, comparative experimental results between our proposed method and the relative state-of-the-art methods will be given.

Table 1: Comparative experimental results on DOTA dataset between different methods.

	mAP	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
FR-O[29]	40.7	58.9	59.2	15.8	36.4	23.9	21.1	28.2	69.0	61.3	50.3	33.0	46.1	37.1	41.3	29.0
RetinaNet[33]	47.6	80.0	53.8	31.8	46.7	46.1	39.4	51.8	90.5	59.6	50.2	26.9	54.7	27.9	27.2	27.5
R2CNN[14]	60.7	88.5	71.2	31.7	59.3	51.9	56.2	57.3	90.8	72.8	67.4	56.7	52.8	53.1	51.9	53.6
RRPN[34]	61.0	80.9	65.8	35.3	67.4	59.9	50.9	55.8	90.7	66.9	72.4	55.1	52.2	55.1	53.4	48.2
LR-O[1]	62.0	81.1	77.1	32.3	72.6	48.5	49.4	50.5	89.9	72.6	73.7	61.4	58.7	54.8	59.0	48.7
DCN[8]	65.0	80.8	77.7	37.2	75.8	58.8	51.1	63.5	88.2	75.5	78.0	57.8	64.0	57.9	59.5	49.7
RT[6]	67.3	81.3	77.5	41.2	72.2	67.9	63.1	72.9	90.8	71.4	76.5	61.4	57.2	66.2	59.1	51.3
ICN[6]	68.2	81.4	74.3	47.7	70.3	64.9	67.8	70.0	90.8	79.1	78.2	53.6	62.9	67.0	64.2	50.2
Mask RCNN[13]	70.3	89.2	76.3	50.8	66.2	**78.2**	75.9	86.1	90.2	81.0	81.9	45.9	57.4	64.8	63.0	47.7
BBAvector[34]	70.9	89.3	81.2	42.3	63.1	75.9	68.3	87.6	**90.9**	85.9	83.6	49.6	62.5	66.3	70.0	48.0
HTC[6]	71.3	89.3	77.0	52.2	66.0	77.9	75.6	86.9	90.5	80.6	80.5	48.7	57.2	69.5	64.6	52.5
R3Det[32]	71.7	89.5	82.0	48.5	62.5	70.5	74.3	77.5	90.8	81.4	83.5	62.0	59.8	65.4	67.5	60.1
FFA[6]	75.7	**90.1**	**82.7**	54.2	75.2	71.0	79.9	83.5	90.7	83.9	84.6	61.2	**68.0**	70.7	**76.0**	**63.7**
ReDet[24]	76.3	88.8	82.6	54.0	74.0	78.1	**84.1**	**88.0**	**90.9**	**87.8**	**85.8**	61.8	60.4	76.0	68.1	63.6
HeatNet	**76.8**	88.7	77.6	**55.3**	**77.2**	78.0	82.8	87.5	90.8	87.3	85.4	**66.0**	63.7	**77.4**	71.6	62.1

4.1 Experimental datasets

DOTA dataset [29], is designed for the task of object detection in remotely sensed images, which is collected from variant sensors and platforms, and consists of 2,806 satellite and aerial images with variant scale, aspect ratio and arbitrary orientation, ranging from 800×800 to 4000×4000. There are 15 categories and 188,282 instances in this dataset. We crop the images into 600×600 patches with a stride of 100 and scales of 0.5 and 1 as in [34].

HRSC2016 dataset [20], is designed for the task of ship detection in remotely sensed satellite images, which is collected from Google Earth, and consists of 1,061 images ranging from 300×300 to 1500×900. This dataset contains 436 training images, 181 validation images, and 444 test images.

4.2 Implement of our proposed HeatNet method

The basic feature extractor is modified by the attention in the frequency channel and followed by 4 prediction branches. The input images are with a dimension of 608×608 and the output heatmaps are 152×152. To extract the center points on a heatmap, NMS is applied through a 3×3 max-pooling layer, top 500 center points being selected from the heatmap. The offset B^O, orientation discrimination B^D and bounding box determination B^B are generated from the center points c_2, which are modified by the offsets $c_1 = c + B^O$ and scaled by the stride $c_2 = sc_1, s = 4$.

Adam with an initial learning rate of 1.25×10^{-5} is selected to optimize the overall loss function $L = L_{B^H} + L_{B^O} + L_{B^D} + L_{B^B}$. We train the network for about 60 epochs on the DOTA dataset and 150 epochs on the HRSC2016 dataset, as usual.

4.3 Compare with other detection methods

To demonstrate the effectiveness of our proposed HeatNet method, 14 representative object detection methods of Object Detection in Aerial Images (ODAI) challenge [11] are compared on the DOTA dataset, the comparative experimental results are shown in Table 1, in which the abbreviation for each category is the same with [34]. From Table 1, we can see that, the object detection accuracy of our proposed HeatNet method surpasses both the anchor based methods and the keypoint based methods, including the latest ReDet method, which has demonstrated the superior of this proposed method.

Table 2: Comparative results on HRSC2016 dataset.

	mAP
CP[20]	55.7
BL2[20]	69.6
RC2[20]	75.7
RT[6]	80.1
BBAvector[34]	82.8
HeatNet	84.7

Table 3: The effectiveness of each part of our proposed HeatNet method.

Combination	att-F	re-F	mAP	△ %
Baseline(BS)			84.4	—-
BS + att-F	√		85.0	0.6%↑
BS + re-F		√	85.5	1.1%↑
BS + HeatNet	√	√	85.9	1.5%↑

We also compare our proposed HeatNet method with other keypoint based detection methods, including the most powerful BBAvector method [34], on the HRSC2016 dataset, the comparative experimental results are shown in Table 2. From Table 2, we can see that, on the HRSC2016 dataset, the detection accuracy of our proposed HeatNet method even surpasses the BBAvector method by 1.9%.

4.4 Ablation studies

In this section, we will validate the effectiveness of each part of our proposed HeatNet method, including the attention-based feature extractor in frequency channel (att-F) and the refinement of FFT-based heatmaps in frequency domain (re-F), on the detection of ship category in the DOTA dataset, and the ablation testing results are as shown in Table 3. From Table 3, we can be seen that, the att-F improves the accuracy of object detection by 0.6%, the re-F improves the accuracy of object detection by 1.1% and the combination of att-F, re-F improves the accuracy of object detection by 1.5%, which has validated the effectiveness of each part of this proposed HeatNet method.

4.5 Attention-based feature extractor in frequency channel

In this experiment, we aim at investigating the effectiveness of the attention-based feature extractor in frequency channel (att-F). We select the BBAvector [34] as the baseline detection method, comparing our proposed method with the baseline on the typical cluster distributed ship category. And we test the proper setting of top k frequency components, which are selected and reserved in the attention-based feature extractor. From the comparative experimental results shown in Table 4, we can see that, the top 16 frequency components contribute most to improving the accuracy of object detection. The reason why the results are uneven for different k is that, the DCT of att-F interacts with the Gaussian distributed center area, resulting from the band-pass effect, and the parameters μ and σ of center area remain unchanged for various k during these experiments.

4.6 Refinement of FFT-based heatmaps in frequency domain

In this experiment, we will verify the effectiveness of refining FFT-based heatmaps in the frequency domain with changing the settings of Eq.(6). We select the BBAvector [34] as the baseline detection method and compare our proposed method on the representative cluster distributed ship category. From the detection results shown in Table 5 and the heatmaps with and without the FFT-based refinement shown in Fig.4, we can see that, the FFT-based

Table 4: Comparative results between with and without the attention-based feature extractor.

Combination	mAP	△ (%)
Baseline(BS)	84 4	—-
BS + att-F:k=8	78.0	6.4%↓
BS + att-F:k=16	85.0	0.6%↑
BS + att-F:k=32	77.4	7.0%↓

Table 5: Comparative results between using and without using heatmap refinement.

Combination	mAP	△ (%)
Baseline(BS)	84.4	—-
BS + re-F:$\alpha_0 = 1, \alpha_h = 0.25, \alpha_l = 0.05$	85.5	1.1%↑
BS + re-F:$\alpha_0 = 1, \alpha_h = 0.25, \alpha_l = 0$	85.3	0.9%↑
BS + re-F:$\alpha_0 = 0.7, \alpha_h = 0.5, \alpha_l = 0$	85.1	0.7%↑
BS + re-F:$\alpha_0 = 0.5, \alpha_h = 0.7, \alpha_l = 0$	85.0	0.6%↑

refinement can relieve the adhesion of cluster distributed targets effectively and a remarkable improvement can be achieved with the refining of FFT-based heatmaps in the frequency domain. In addition, experimental results show that, when the sum of α_0, α_h and α_l is larger than 1.25, the network will be difficult to converge. Therefore, the varying range of the sum of α_0, α_h and α_l is limited to less than 1.25. Among these parameters, the most proper setting of Eq.(6) is $\alpha_0 = 1, \alpha_h = 0.25, \alpha_l = 0.05$.

5 Conclusion

For the detection of cluster distributed targets in satellite images, a refinement of FFT-based heatmap with multi-branches network is proposed, which includes the refinement of FFT-based heatmaps in frequency domain and an attention-based feature extractor in frequency channel, to focus the attention and refine the salient regions for the cluster distributed targets. Additionally, a keypoint-based detection is adopted as the basic workflow to tackle with the adhesion, a scale-aware center area is conducted to tackle with the variation of scale, and an orientation discrimination is also utilized to eliminate the specificity of different targets. To our best knowledge, this is the first time that the common impact factors of remote sensing object detection have been formulated quantitatively. And this is the first time the cluster distribution of remotely sensed targets has been studied and associated this task in the frequency domain, and the FFT is utilized to refine heatmap rather than accelerating CNN calculation. The effectiveness of our proposed method has been validated on two public datasets, and the comparative experimental results with different state-of-the-art methods have demonstrated the superior of this proposed method.

6 Acknowledgement

This work was supported in part by the National Nature Science Foundation (41971294), China Postdoctoral Science Foundation (2020M680560) and Cross-Media Intelligent Technology Project of BNRist (BNR2019TD01022) of China.

References

[1] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, pages 150–165. Springer, 2018.

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.

[3] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing convolutional neural networks in the frequency domain. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1475–1484, 2016.

[4] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, David Belanger, Lucy Colwell, et al. Masked language modeling for proteins via linearly scalable long-context transformers. *arXiv preprint arXiv:2006.03555*, 2020.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[6] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019.

[7] Max Ehrlich and Larry S Davis. Deep residual learning in the jpeg transform domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3484–3493, 2019.

[8] Hazem M El-Bakry and Qiangfu Zhao. Fast object/face detection using neural networks and fast fourier transform. *International Journal of Signal Processing*, 1(3):182–187, 2004.

[9] Kun Fu, Zhonghan Chang, Yue Zhang, Guangluan Xu, Keshu Zhang, and Xian Sun. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 161: 294–308, 2020.

[10] Himanshu Gothwal, Silky Kedawat, Rajesh Kumar, et al. Cardiac arrhythmias detection in an ecg beat signal using fast fourier transform and artificial neural network. *Journal of Biomedical Science and Engineering*, 4(04):289, 2011.

[11] Xiang Bai Gui-Song Xia. Object detection in aerial images (odai). https://captain-whu.github.io/ODAI/index.html.

[12] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[14] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.

[15] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[16] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021.

[17] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*, 2017.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[19] Zhenhua Liu, Jizheng Xu, Xiulian Peng, and Ruiqin Xiong. Frequency-domain dynamic pruning for convolutional neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1051–1061, 2018.

[20] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods*, volume 2, pages 324–331. SCITEPRESS, 2017.

[21] Zili Liu, Tu Zheng, Guodong Xu, Zheng Yang, Haifeng Liu, and Deng Cai. Training-time-friendly network for real-time object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11685–11692, 2020.

[22] Michael Mathieu, Mikael Henaff, and Yann LeCun. Fast training of convolutional networks through ffts. *arXiv preprint arXiv:1312.5851*, 2013.

[23] Kei-ichiro Minami, Hiroshi Nakajima, and Takeshi Toyoshima. Real-time discrimination of ventricular tachyarrhythmia with fourier-transform neural network. *IEEE transactions on Biomedical Engineering*, 46(2):179–185, 1999.

[24] Martina Mironovova and Jirí Bíla. Fast fourier transform for feature extraction and neural network for classification of electrocardiogram signals. In *2015 Fourth International Conference on Future Generation Communication Technology (FGCT)*, pages 1–6. IEEE, 2015.

[25] Harry Pratt, Bryan Williams, Frans Coenen, and Yalin Zheng. Fcnn: Fourier convolutional neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 786–798. Springer, 2017.

[26] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. *arXiv preprint arXiv:2012.11879*, 2020.

[27] Jinwang Wang, Wen Yang, Heng-Chao Li, Haijian Zhang, and Gui-Song Xia. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4307–4323, 2020.

[28] Haoran Wei, Yue Zhang, Zhonghan Chang, Hao Li, Hongqi Wang, and Xian Sun. Oriented objects as pairs of middle lines. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:268–279, 2020.

[29] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.

[30] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749, 2020.

[31] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[32] Xue Yang, Qingqing Liu, Junchi Yan, Ang Li, Zhiqiang Zhang, and Gang Yu. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612*, 2019.

[33] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8232–8241, 2019.

[34] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris Metaxas. Oriented object detection in aerial images with box boundary-aware vectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2150–2159, 2021.

[35] Jiong Zhang, Yibo Lin, Zhao Song, and Inderjit Dhillon. Learning long term dependencies via fourier recurrent units. In *International Conference on Machine Learning*, pages 5815–5823. PMLR, 2018.

[36] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1745–1749, 2018.

[37] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[38] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 850–859, 2019.