Received: 26 June 2023 | Revised: 10 October 2023 | Accepted: 6 November 2023

DOI: 10.1002/spe.3287

SPECIAL ISSUE PAPER

WILEY

Object counting in remote sensing via selective spatial-frequency pyramid network

Jinyong Chen[1] | **Mingliang Gao**[1] | **Xiangyu Guo**[1] | **Wenzhe Zhai**[1] | **Qilei Li**[2] | **Gwanggil Jeon**[3]

[1]School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China

[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

[3]Department of Embedded Systems Engineering, Incheon National University, Incheon, South Korea

Correspondence
Mingliang Gao, School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China.
Email: mlgao@sdut.edu.cn
Gwanggil Jeon, Department of Embedded Systems Engineering, Incheon National University, Incheon, South Korea.
Email: gjeon@inu.ac.kr

Abstract

The integration of remote sensing object counting in the Mobile Edge Computing (MEC) environment is of crucial significance and practical value. However, the presence of significant background interference in remote sensing images poses a challenge to accurate object counting, as the results are easily affected by background noise. Additionally, scale variation within remote sensing images presents a further difficulty, as traditional counting methods face challenges in adapting to objects of different scales. To address these challenges, we propose a selective spatial-frequency pyramid network (SSFPNet). Specifically, the SSFPNet consists of two core modules, namely the pyramid attention (PA) module and the hybrid feature pyramid (HFP) module. The PA module accurately extracts target regions and eliminates background interference by operating on four parallel branches. This enables more precise object counting. The HFP module is introduced to fuse spatial and frequency domain information, leveraging scale information from different domains for object counting, so as to improve the accuracy and robustness of counting. Experimental results on RSOC, CARPK, and PUCPR+ benchmark datasets demonstrate that the SSFPNet achieves state-of-the-art performance in terms of accuracy and robustness.

KEYWORDS

attention mechanism, background clutter, edge computing, object counting, remote sensing, scale variation

1 | INTRODUCTION

Mobile Edge Computing (MEC) is an emerging computing model characterized by real-time capabilities, low latency, and security features, which drives the development of technology and applications in fields such as the Internet

Abbreviations: AI, artificial intelligence; CARPK, car parking lot; DCT, discrete cosine transform; FFB, frequency feature branch; HFP, hybrid feature pyramid; IDCT, inverse discrete cosine transform; MACs, multiply-accumulates; MAE, mean absolute error; MEC, mobile edge computing; MSE, mean squared error; Params, parameters; PA, pyramid attention; PCA, pyramid channel attention; PGC, pyramid grouped convolution; PUCPR+, Pontifical Catholic University of Parana+; PSA, pyramid spatial attention; RMSE, root mean square error; RSOC, remote sensing object counting; SFB, spatial feature branch; SOTA, state-of-the-art; SSFPNet, selective spatial-frequency pyramid network.

of Things and smart cities.[1-3] In the MEC environment, the integration of remote sensing object counting holds significant importance and practical value. Remote sensing object counting, as a specific Artificial Intelligence (AI) application, leverages the computing capabilities of edge servers to achieve real-time and accurate counting of objects in remote sensing imagery. This integration plays a crucial role in addressing the challenges faced by AI applications in the MEC environment.[4] The integration of remote sensing object counting with the MEC environment brings numerous benefits. Firstly, by performing object counting at the edge, the computational burden on centralized cloud servers is reduced, which improves efficiency and responsiveness.[5] This allows for real-time processing of remote sensing imagery, which is particularly critical in applications requiring timely and accurate information, such as environment monitoring,[6-8] urban planning,[9,10] and disaster management.[11] Secondly, the integration enhances data security and privacy protection. By processing data at the edge servers, the need for transmitting raw data is minimized, which reduces the risk of data breaches and preserves personal privacy.[5] Lastly, the combination of remote sensing object counting with the MEC environment enables the deployment of AI-based applications that rely on accurate object counts. This opens up a wide range of possibilities and applications in various domains, which contributes to advancements in environmental monitoring, urban planning, disaster management, and more.

Currently, object counting methods can be divided into three categories, that is, detection-based,[12,13] regression-based,[14] and density map estimation based methods.[15] Detection-based algorithms commonly employ detection models to identify and localize objects within an image. It estimates the overall count of target objects by leveraging the number of detected instances. Dollar et al.[16] used a sliding window on the images to detect the number of objects. For complex background and occlusion phenomena, the effect of the method is poor. To alleviate these challenges, regression-based methods have been introduced, which directly learn the mapping between object features and the number of objects. Regression models such as support vector regression and random forest regression[17] are used to perform feature extraction and quantity prediction on input images. Despite its proficiency in mitigating challenges like occlusion and background clutter,[18] these methods frequently neglect spatial information and tend to extract low-level information. It is hard to regress a high-quality density map. The density estimation approach utilizes a pretrained deep learning model to process the input image and generate a density map.[19] By summing the pixel values in the density map, the total count of objects can be determined. For example, Guo et al.[20] proposed a dense attention fusion network for counting objects in remote sensing imagery. This method takes advantage of a pretrained model to estimate the density of objects in the image and uses the accumulated density values for estimating the overall object count.

Although density map estimation methods have become mainstream for object counting tasks, counting objects in remote sensing imagery poses background clutter and scale variation. As shown in Figure 1, the red box in the first column represents the presence of background interference, while the second column displays the scale variations. Unlike object counting in common camera monitoring, the objects captured by drones often vary significantly in size, shape, and appearance, making them difficult to detect and distinguish from the background. To address the background noise, researchers have developed segmentation-based[21-24] and attention-based[25,26] methods for object counting in remote sensing. Segmentation techniques aid in distinguishing between target and background regions. Cholakkal et al.[27] proposed a method that utilizes segmentation techniques to segment the target regions within the images, which results in generating segmentation masks for each target instance. As object counting and instance segmentation require detailed pixel-level annotations, they necessitate consideration of data and computational costs. Take the above issue into consideration, the attention mechanism has been extensively employed to suppress background clutter and emphasize foreground regions. Guo et al.[28] proposed a triple view attention module to compensate for the impact of background noise, and distinguish object regions by performing three-dimensional operations to capture the interactive dependencies between them. Zhai et al.[26] put forward a channel attention module that effectively reduces misestimations in background regions by selectively attending to important channel features. The scale variation arises due to the wide range of object sizes present, ranging from small objects covering just a few pixels to large objects spanning thousands of pixels. To address the scale variation, researchers are devoted to multiscale feature fusion approaches[29,30] that analyze remote sensing imagery at different resolutions or pyramid levels. Multiscale analysis helps capture objects that may be missed or inaccurately counted at a single scale. Gao et al.[11] adopted a scale pyramid with various dilation rates to capture multiscale information in remote sensing images. Different from this, Yu et al.[31] extracted frequency domain features through mathematical transformations and constructed feature pyramids based on preserving different low-frequency information to capture multiscale information. By contrast, the former

FIGURE 1 Exemplars of objects exhibiting varying scales and background clutter in remote sensing images.

ignores the scale information in the frequency domain, while the latter ignores the scale information in the spatial domain.

Therefore, we combine feature information in both spatial and frequency domains and propose a Selective Spatial-Frequency Pyramid Network (SSFPNet). The proposed SSFPNet consists of two modules, that is, the pyramid attention (PA) module, and the hybrid feature pyramid (HFP) module. The PA module begins by employing grouped convolutions with varying configurations to generate a distinct feature pyramid. Subsequently, spatial and channel attention modules are successively applied to the four branches. These modules effectively suppress irrelevant interference regions and channels, improving the understanding of target information while reducing sensitivity to the background. The HFP module comprises two types of branches: a spatial branch that captures features with different receptive fields, and a frequency branch that captures features with different low-frequency energy characteristics. These branches are integrated to form a multiscale pyramid, effectively handling the significant scale variations. In summary, the contributions of this work are as follows.

1. A PA module is introduced to tackle the background interference. The pyramid spatial attention (PSA) and pyramid channel attention (PCA) in the PA module, are introduced to extract abundant feature information without increasing the amount of calculation.
2. An HFP module is constructed to address scale variation. It includes a spatial feature branch and three frequency feature branches. The spatial feature branch is to extract large-scale information. In contrast, the frequency feature branch mainly focuses on capturing small-scale information present in the frequency domain.
3. Extensive experimental results on RSOC dataset, CARPK, and PUCPR+ demonstrate the accuracy and robustness of SSFPNet. Furthermore, the effectiveness of the proposed approach, which combines the spatial and frequency domains, was validated by the ablation experiments.

The rest of the paper is organized as follows. In Section 2, some research works related to this paper are reviewed. Section 3 provides a detailed description of the proposed method. Detailed experimental analyses are presented in Section 4. Section 5 provides the conclusion and future work.

2 | RELATED WORK

2.1 | Object counting in spatial domain

The most common methods in the spatial domain are based on multiscale approaches and attention mechanisms. The multiscale method, first introduced in Reference 32, addressed the issue of multiple scales by using three-column CNNs with different-sized convolutional kernels. Each column network captures features at a special scale, enabling the model to handle objects of diverse sizes and scales. Training the network on multiple scales, it enhances the network's ability to learn from different scale information. Another way to solve the scale issue is to use dilated convolution. Chen et al.[33] proposed a scale pyramid with diverse receptive fields, which employs different dilation rates to capture features in various scales while keeping the output resolution unchanged. Nevertheless, the adoption of a multi-column structure introduces a considerable amount of information redundancy and leads to an increase in the model's complexity. To address this drawback, Li et al.[34] proposed a single-column network architecture with a cascade expansion function to extract multiscale information. Although the method based on multiscale CNNs can be used to deal with scale variation, it ignores the abundant spatial location information. Meanwhile, when the attention mechanism is introduced into the field of object counting, rich context information can be extracted. Gao et al.[35] proposed regression networks with spatial and channel-wise attention to mitigate the interference of the background through the attention module. Zhai et al.[36] introduced the use of a dual attention module, incorporating both channel and spatial attention, to enhance counting accuracy. Jiang et al.[37] proposed an attention scaling factor to assign different attention masks to different density levels, which helps to weaken the estimation error in different regions. To address both background and scale issues, Zhai et al.[38] proposed a novel attention hierarchy ConvNet, which includes a recalibrated attention module to suppress background interference and a feature enhancement module with varying scales. Therefore, CNNs methods combining multiscale information and attention mechanisms are a mainstream approach in object counting. Similarly, we also utilize attention mechanisms to attenuate the influence of background interference in the spatial and channel domain.

2.2 | Object counting in frequency domain

Compared with the traditional spatial methods, the feature map captured in the frequency domain has more contextual information than the feature map generated by spatial domain.[31] Idrees et al.[39] achieved more accurate count estimations in complex scenarios by considering the confidence associated with observing individuals and utilizing frequency-based analysis. Ehrlich et al.[40] introduced frequency domain transformation in CNNs. The transformed coefficients are quantized and encoded, while also incorporating deep residual learning in the frequency domain. Xu et al.[41] added DCT transform to the neural network to reduce the bandwidth, which has high accuracy compared with the traditional spatial downsampling. Qin et al.[42] proved that the commonly used global average pooling can be viewed as a specific instance of feature decomposition in the frequency domain. Shu et al.[43] proposed an efficient characteristic function loss by transforming the density map to the frequency domain and the properties of the characteristic. The advantage of the characteristic function loss function is that no external algorithm is needed to extract spatial information. Guo et al.[44] proposed a spatial-frequency attention network, which combines spatial location information and frequency-domain characteristic information to solve the issue of large-scale variation in dense scenes. Different from the previous work, we combine the scale information of different receptive fields in the spatial domain and the low-frequency scale information with different coefficients in the frequency domain to extract multiscale information.

3 | METHODOLOGY

3.1 | Architecture overview

To tackle the background interference and scale variation in remote sensing object counting, we propose the SSFPNet, which is illustrated in Figure 2. It consists of four modules, that is, a backbone module, a PA module, an HFP module, and a backend module. First, the VGG16 serves as the backbone to extract low-level features. Next, the PA module is built to highlight discriminative features and reduce sensitivity to the background. Then, multiscale features are extracted

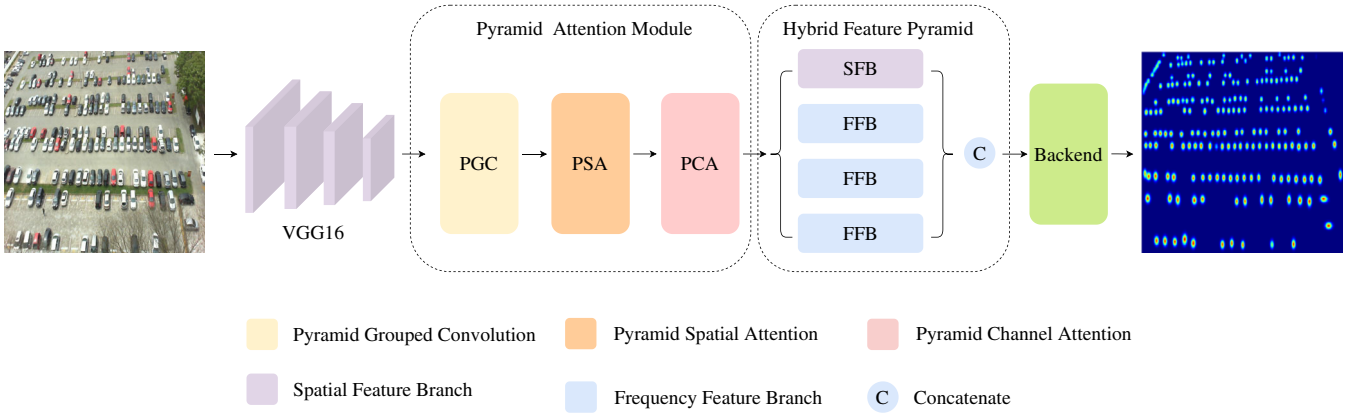

FIGURE 2 Framework of the proposed SSFPNet for object counting.

through the HFP module composed of spatial and frequency domain pyramids. Finally, We deployed a series of convolutional layers in the backend to aggregate scale information from different domains. The backend module aims to integrate the spatial and frequency domain features, leveraging their complementary characteristics. This aggregation enhances the model's ability to perceive and understand the structural and scale variations in remote sensing images, leading to improved detection accuracy for objects of different scales. To generate a density map, the module utilizes a 1×1 convolutional layer, which integrates information from the preceding layers.

3.2 | Pyramid attention module

Counting objects in remote sensing is susceptible to the influence of background noise. Therefore, a PA module is introduced to address the background noise by capturing spatial relationships between objects and backgrounds. The PA module comprises three units, that is, a pyramid grouped convolution (PGC) unit, a pyramid spatial attention (PSA) unit, and a pyramid channel attention (PCA) unit. These units collaborate to improve the attention mechanism of the model and capture comprehensive information. By employing the PGC unit, the module performs grouped convolutions with varying group sizes across multiple branches. This increases the receptive field for each group and thus enables the model to better capture spatial relationships between objects and backgrounds. Each group can focus on learning features from specific regions, which enhances the model sensitivity to local features of both the targets and the background. The PSA unit enables the model to focus on important spatial regions while suppressing background noise, thereby improving the precision and accuracy of object counting. Additionally, the PCA unit enhances the representation of channel-wise features, enabling the model to extract and emphasize key information for counting remote-sensing objects.

Pyramid grouped convolution unit: In order to obtain more location information, four branches are deployed to extract spatial information with different scales. The group convolution method is introduced to deal with four group features of diverse nuclear scales without increasing the amount of computation. The framework of the PGC unit is shown in Figure 3. To enhance the location information, four branches are utilized to extract spatial features. The group convolution technique is employed to process the four groups of features with distinct receptive fields without significantly increasing the computational complexity. This approach allows for capturing diverse spatial details while maintaining computational efficiency. The framework of the Pyramid Grouped Convolution (PGC) unit can be seen in Figure 3, showcasing the integration of four branches and the group convolution operation. The low-level features obtained from the front end are fed into four parallel branches, each utilizing different convolutional kernel sizes and group convolution numbers. By employing this approach, the model can effectively harness the multiscale information contained in the low-level features. The correlation between the size of the multiscale kernel and the group size is as follows,

$$g_i = 2^{\frac{k_i-1}{2}}, \quad k_i = 2 \times (i+1) + 1, \quad i = 0, 1, 2, 3 \tag{1}$$

where g_i denotes the number of groups in the ith branch. k_i indicates the convolution kernel size of the ith branch.

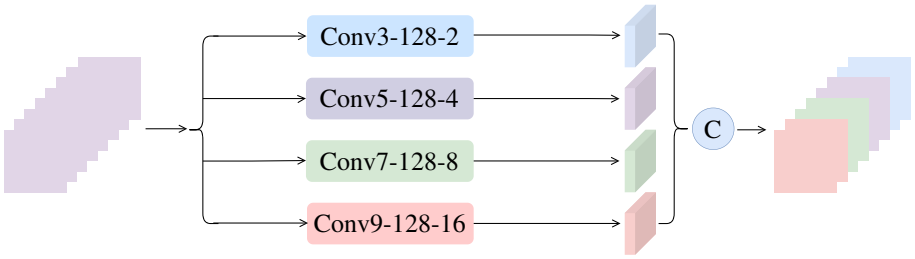

FIGURE 3 Framework of the pyramid grouped convolution (PGC) unit. $Conv(k) - (c) - (g)$ represents the group convolution operation, where k, c, and g represent the kernel size, the number of input channels, and groups, respectively.

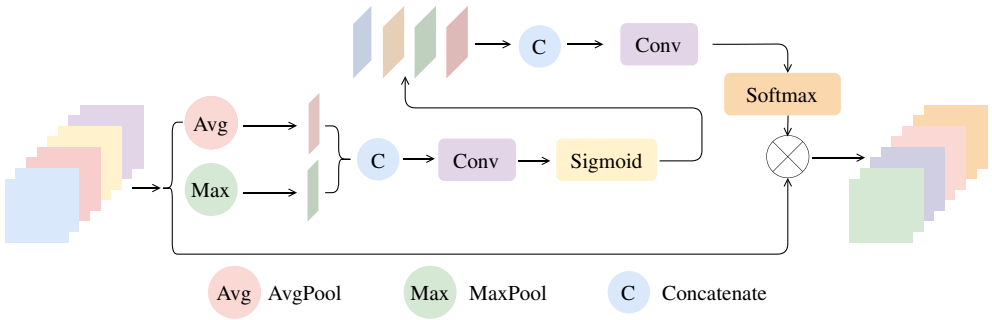

FIGURE 4 Structure of the pyramid spatial attention unit.

Pyramid spatial attention unit: The structure of the PSA unit is shown in Figure 4. Attention weights are learned separately on the four branches, effectively reducing the model's sensitivity to the background and highlighting spatially significant features. This process enables the extraction of valuable spatial information from the input data, enhancing the model's capacity to concentrate on relevant details. Then, the information is integrated through the concatenate operation to combine the local details with the global context. This allows the model to capture both the details and local features of the targets simultaneously. Thus, it can enhance the accuracy and robustness of object counting. Finally, the Softmax operation was used to help the model focus on the target area of interest.

By extracting spatial attention weight from feature maps of four branches, the attention weight vector can be formulated as,

$$Y_i = Sigmoid(Conv([MaxPool(X_i); AvgPool(X_i)]), \quad i = 0, 1, 2, 3, \tag{2}$$

Where Y_i is the spatial attention weight of the ith branch, respectively. $Maxpool(\cdot)$ and $Avgpool(\cdot)$ denote average pooling and max pooling along the channel dimension, respectively. $Conv(\cdot)$ indicates the dimension reduction operation. By adjusting the local spatial attention maps Y_i corresponding to the four branches, the global spatial attention can be recalibrated. In a nutshell, the PSA unit is formulated as follows,

$$F_s = softmax(Conv(Cat(Y_0, Y_1, Y_2, Y_3))), \tag{3}$$

where $Cat(\cdot)$ denotes the concatenate operation. F_s represents the learned spatial attention weight.

Pyramid channel attention unit: The structure of the pyramid channel attention unit is shown in Figure 5. The significance of distinctive channels within each group is individually constructed using the PCA unit. The channel weights of the four branches are dynamically adjusted. Then, the channel weight of the whole feature map is reassigned through concatenation and Softmax operation. Each branch of channel attention focuses on capturing and highlighting different aspects of the features related to the objects of interest. The PCA mechanism applies attention weights to different channel groups, enabling the model to focus on the most informative and discriminative channels. By leveraging channel

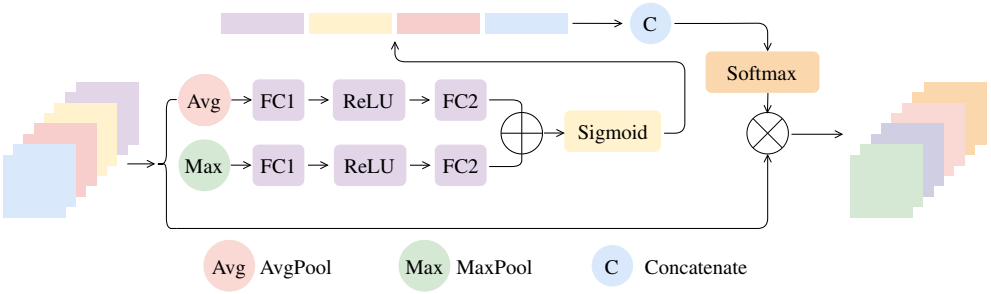

FIGURE 5 Structure of pyramid channel attention unit.

attention in multiple branches, the model can effectively extract and utilize the most relevant features, ultimately improving its performance in remote sensing object counting tasks.

Assuming the input feature map is denoted as $X \in \mathbb{R}^{C\times H\times W}$, the PCA unit applies average pooling and max pooling operations to each channel of the feature map $X_c \in \mathbb{R}^{H\times W}$. The formulas for these operations are as follows:

$$\begin{aligned} MaxPooling_c &= max(X_c), \\ AvgPooling_c &= \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X_c(i,j), \end{aligned} \tag{4}$$

where X_c denotes the cth channel of the feature map X. $max(\cdot)$ denotes the operation that takes the maximum values. $X_c(i,j)$ denotes the element value of the feature map X at channel c, height i, width j. In this way, for each way, c, a maximum value, and an average value are obtained as the pooling results for the channel. The description of channel attention for each branch is given below.

$$Y_i = Sigmoid(f_{c2}(f_{c1}(MaxPooling_c)) + f_{c2}(f_{c1}(AvgPooling_c))), \quad i = 0, 1, 2, 3, \tag{5}$$

where Y_i indicates the channel attention weight of the ith branch. f_{c1} and f_{c2} denote the fully connected layer. The attention weight of the whole channel is obtained by concatenating the four channel branches. It is formulated as:

$$F_c = Softmax(Cat(Y_0, Y_1, Y_2, Y_3)), \tag{6}$$

where $Cat(\cdot)$ denotes the operation of concatenate. F_c is the attention weight of the whole channel.

3.3 | Hybrid feature pyramid module

Constructing a spatial feature pyramid is a commonly used method for extracting multiscale information. However, to compensate for the limitations of scale information in the spatial domain, we introduce three branches to extract frequency domain feature information. In the frequency domain, we utilize Discrete Cosine Transform (DCT) and Inverse Discrete Cosine Transform (IDCT) to extract multiscale features. By preserving different DCT coefficients, we can construct frequency domain branches with different information, forming a frequency domain pyramid. In the frequency feature branch (FFB), we utilize DCT to construct frequency domain feature information. The 3D DCT transform is formulated as,

$$F(u,v,w) = c(u)c(v)c(w)\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\sum_{k=0}^{C-1}f(i,j,k)cos\frac{(i+0.5)u\pi}{H}cos\frac{(j+0.5)v\pi}{W}cos\frac{(k+0.5)w\pi}{C}, \tag{7}$$

where u, v, and w represent frequency domain variables transformed in x, y and z dimensions, while $c(u), c(v)$ and $c(w)$ denote constant terms. Let $c(u) = c(v) = c(w)$, where $c(u)$ has the following expression,

$$c(u) = \begin{cases} \sqrt{\frac{1}{H}}, & \text{if } u = 0, \\ \sqrt{\frac{2}{H}}, & \text{otherwise.} \end{cases} \tag{8}$$

The discrete inverse cosine transform (IDCT) formula is as follows,

$$f(i,j,k) = c(u)c(v)c(w)\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\sum_{w=0}^{C-1}F(u,v,w)cos\frac{(i+0.5)u\pi}{H}cos\frac{(j+0.5)u\pi}{W}cos\frac{(z+0.5)w\pi}{C}. \tag{9}$$

The HFP module consists of four branches, as shown in Figure 6. It consists of one spatial feature branch (SFB) and three frequency feature branches (FFB). The SFB utilizes dilated convolutions with a single dilation rate primarily to extract large-scale information in the spatial domain, which compensates for the spatial scale information missing in the frequency domain. The FFB retains 1/16, 1/64, and 1/256 of the DCT coefficients respectively, to capture small-scale information in the frequency domain.

Dilated convolutions expand the receptive field and improve the model's responsiveness to objects of different scales. This enables the model to handle objects at various scales more accurately and robustly. Frequency domain features are crucial for describing image textures and details. Therefore, the introduction of frequency domain branches enhances the model's perception of image details. By preserving low-frequency information at multiple scales in the frequency domain, a more comprehensive representation of the image's structure and scale features is obtained. This allows the model to better understand the global context and overall characteristics of the image. Combining the scale features from the spatial branch with the frequency domain information enables the model to leverage the advantages of both spatial and frequency domain features. Thus, it provides a more comprehensive description of the image's scale variations and frequency characteristics. Additionally, this approach improves the counting performance of the model, enabling it to handle diverse scales and frequency domain features in remote sensing image scenes.

The backend module is illustrated in Figure 6. It is responsible for aggregating the spatial and frequency domain information to achieve accurate remote sensing object counting. The deformable convolution is to deal with arbitrary orientations in remote sensing images. The primary function of this module is to integrate the spatial and frequency domain features. By this means, it captures the most informative aspects of these features and reduces their dimensionality. This compression and representation process enables the module to focus on the essential and discriminative characteristics of the features, and it can enhance the overall effectiveness of the aggregation process. The backend module serves as a vital component for aggregating spatial and frequency domain information. Through its integration of convolutional structures, it effectively combines and utilizes the most relevant and discriminative features.

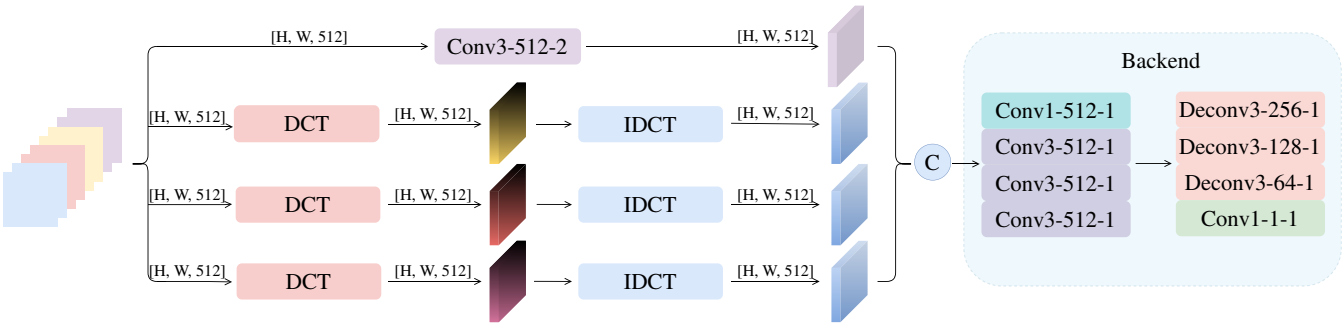

FIGURE 6 Framework of the hybrid feature pyramid (HFP) module and the backend. *Conv*$(k) - (c) - (d)$ refers to ordinary convolution and aims to extract high-level semantic features. *Deconv*$(k) - (c) - (d)$ denotes deformable convolution. (k, c, and d denote the kernel size, the number of output channels, and the dilated rate, respectively. $[H, W, 512]$ in which H, W, and 512 represent the height, width, and number of channels of the feature maps, respectively.) Finally, a 1×1 convolution layer is used to generate the density map.

3.4 | Ground truth generation

The most common way for ground truth generation is applying Gaussian kernel blurred dot annotation.[25,45] In order to enable the image to be processed by Gaussian filtering, a given remote sensing image $I(x)$ needs to be redefined using the impulse function $\delta(x)$, which is expressed as,

$$I(x) = \sum_{i=1}^{N} \delta(x - x_i), \tag{10}$$

where N is the number of objects marked in the image, and x_i is the coordinates of marked points.

Since the generated density map is not continuous, the kernel of a Gaussian filter is applied to each pixel on the image, and the image is smoothed by a convolution operation. Finally, the density map is generated by the following formula,

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x), \tag{11}$$

where the standard deviation σ determines the smoothness of the filter, which is the main parameter of the Gaussian filter.

3.5 | Loss function

The Euclidean function has been utilized to quantify the error between the predicted density map and the ground truth.[11,32] The Mean Squared Error (MSE) loss function is formulated as follows,

$$loss = \frac{1}{N} \sum_{i=1}^{N} \|F_\theta(I_i) - Y_i\|_2^2, \tag{12}$$

where N denote the batch size and θ denotes the learnable parameters. Y_i represents the ground truth, while $F_\theta(I_i)$ represents the estimated density map.

4 | EXPERIMENTS

4.1 | Datasets

Experiments are conducted on remote sensing object counting dataset (RSOC),[11] car parking lot dataset (CARPK),[46] and Pontifical Catholic University of Parana+ Dataset (PUCPR+).[46]

The RSOC dataset[11] is the largest and most commonly used remote sensing counting dataset, which consists of 3057 satellite images with 286,539 annotations. It is divided into four subsets, that is, building, small-vehicle, large-vehicle, and ship.

The CARPK dataset[46] comprises drone-view images obtained from four distinct parking lots, featuring a total of 1448 images. These images are accompanied by 89,777 annotations, indicating the presence of vehicles. It is split into a training set, consisting of 989 images, and a test set, comprising 459 images.

The PUCPR+ dataset[46] is a large-scale vehicle counting dataset containing different weather environments. The dataset contains 125 images with a total of 16,456 annotations. For training purposes, 100 images are utilized, while the remaining 25 images are used for testing. A detailed description and some exemplars of these datasets are shown in Table 1 and Figure 7.

4.2 | Implementation details

All the experiments in this study are conducted using the PyTorch framework. The training and testing processes are performed on an NVIDIA RTX3080Ti GPU. The weight parameters of the trained model are adjusted using the Adam

TABLE 1 Detailed information of the RSOC, CARPK and PUCPR+ datasets.

Datasets	Platform	Images	Train/Test	Size(Avg.)	Annotation Format
RSOC_Building	Satellite	2468	1205/1263	512×512	Center point
RSOC_Large-vehicle	Satellite	172	108/64	1552×1573	Bounding box
RSOC_Small-vehicle	Satellite	280	222/58	2473×2339	Bounding box
RSOC_Ship	Satellite	137	97/40	2558×2668	Bounding box
CARPK	Drone	1448	989/459	720×1280	Bounding box
PUCPR+	Camera	125	100/25	720×1280	Bounding box

FIGURE 7 Exemplars of object counting datasets.

optimizer, with the initial learning rate set to 1e−7 and the weight decay set to 5e−4. The image size small-vehicle, large-vehicle, and ship subdatasets are resized to 1024 × 768 to reduce GPU memory usage due to the large resolution. To enhance the diversity of the dataset and increase the training data volume, we employed the technique of random cropping the images to one-fourth of their original size. This augmentation strategy facilitates the model in better adapting to targets at different positions and scales during the training process. For the small-vehicle, large-vehicle, and ship datasets, it is necessary to convert the bounding box annotations into center-point annotations. We have followed the method presented in Reference 11 to convert the bounding box annotations into center-point annotations. Subsequently, the ground truth can be generated based on the approach illustrated in Section 3.4.

4.3 | Evaluation protocols

The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are adopted to measure the accuracy and robustness of the model. They are formulated as,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{13}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2}, \tag{14}$$

where N is the number of test images. y_i and $\hat{y}_i$ represent the predicted value and the ground truth of the ith image, respectively.

4.4 | Experiments on RSOC dataset

Table 2 provides a comprehensive comparison between the proposed method and other state-of-the-art (SOTA) approaches on the RSOC dataset.[11] These experimental results demonstrate that the proposed method effectively captures more scale information, reduces sensitivity to background noise, and achieves high accuracy and robustness. Compared with ASPDNet,[11] that solely relies on spatial pyramids, the SSFPNet shows significant improvements in MAE and RMSE with a boost of 62.03% and 69.70%, respectively, on the ship dataset. Compared with TSANet,[28] which utilizes the Trip View Attention (TVA) module, the SSFPNet exhibits significant improvements in MAE and RMSE on the building dataset, with an improvement of 10.75% and 11.82% respectively, on the building dataset.

Meanwhile, the results demonstrate that the proposed SSFPNet achieves an MAE of 230.22 and an MSE of 774.77 on the small-vehicle dataset, both ranking first among the compared methods. Compared to CSRNet[34] and ASPDNet[11] that utilize dilated convolutions, the proposed SSFPNet reduces the MAE by 48.13% and 46.86% and the RMSE by 38.13% and 37.45%, respectively. Compared to SFANet[51] and SPNet,[33] the proposed SSFPNet achieves a reduction of 31.06% and 44.71% in MAE and 38.89% and 43.28% in RMSE, respectively, on Large vehicle dataset. These findings demonstrate the effectiveness of SSFPNet in addressing the scale variation problem in object counting tasks.

The visualization results for the four subdatasets of RSOC[11] are shown in Figures 8, 9, 10, and 11, respectively. The top row represents the original remote sensing images, the middle row displays the ground truth, and the bottom row shows the predicted density maps. It can be observed that the predicted results are able to clearly express the density variations of the targets and accurately reflect the level of aggregation and distribution patterns in different areas. On the building dataset with large-size objects, it can be visually observed that the overall density distribution is similar to the ground truth. For the small-vehicle dataset, where the relatively small objects are significantly influenced by the background, it can be observed that the performance is slightly worse at the edges. This proves that the pyramid attention module indeed mitigates the model's sensitivity to background interference. For the large vehicle dataset, the predicted results are able to clearly reflect the positions of the small cars, and the predicted counts are also relatively close to the ground truth. From the predicted ship dataset, it can be observed that SSFPNet effectively handles scale variations, which demonstrates the effectiveness of the hybrid feature pyramid module.

TABLE 2 Comparison results on the RSOC dataset.

	Building		Small vehicle		Large vehicle		Ship	
Methods	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
MCNN[32]	13.65	16.56	488.65	1317.44	36.56	55.55	263.91	412.30
CMTL[47]	12.78	15.99	490.53	1321.11	61.02	78.25	251.17	403.07
CSRNet[34]	8.00	11.78	443.81	1252.22	34.10	46.42	240.01	394.81
SANet[48]	29.01	32.96	497.22	1276.66	62.78	79.65	302.37	436.91
SFCN[49]	8.94	12.87	440.70	1248.27	33.93	49.74	240.16	394.81
SPN[33]	7.74	11.48	445.16	1252.92	36.21	50.65	241.43	392.88
SCAR[35]	26.90	31.35	497.22	1276.65	62.78	79.46	302.37	436.92
CAN[50]	9.12	13.38	457.36	1260.39	34.56	49.63	282.69	423.44
SFANet[51]	8.18	11.75	435.29	1284.15	29.04	47.01	201.61	332.87
ASPDNet[11]	7.59	10.66	433.23	1238.61	31.76	40.14	193.83	318.95
TASNet[28]	7.63	11.25	394.89	1196.83	22.75	37.13	191.82	278.17
SSFPNet(Ours)	**6.81**	**9.92**	**230.22**	**774.77**	**20.02**	**28.73**	**73.59**	**96.64**

Note: The best results are highlighted in bold.

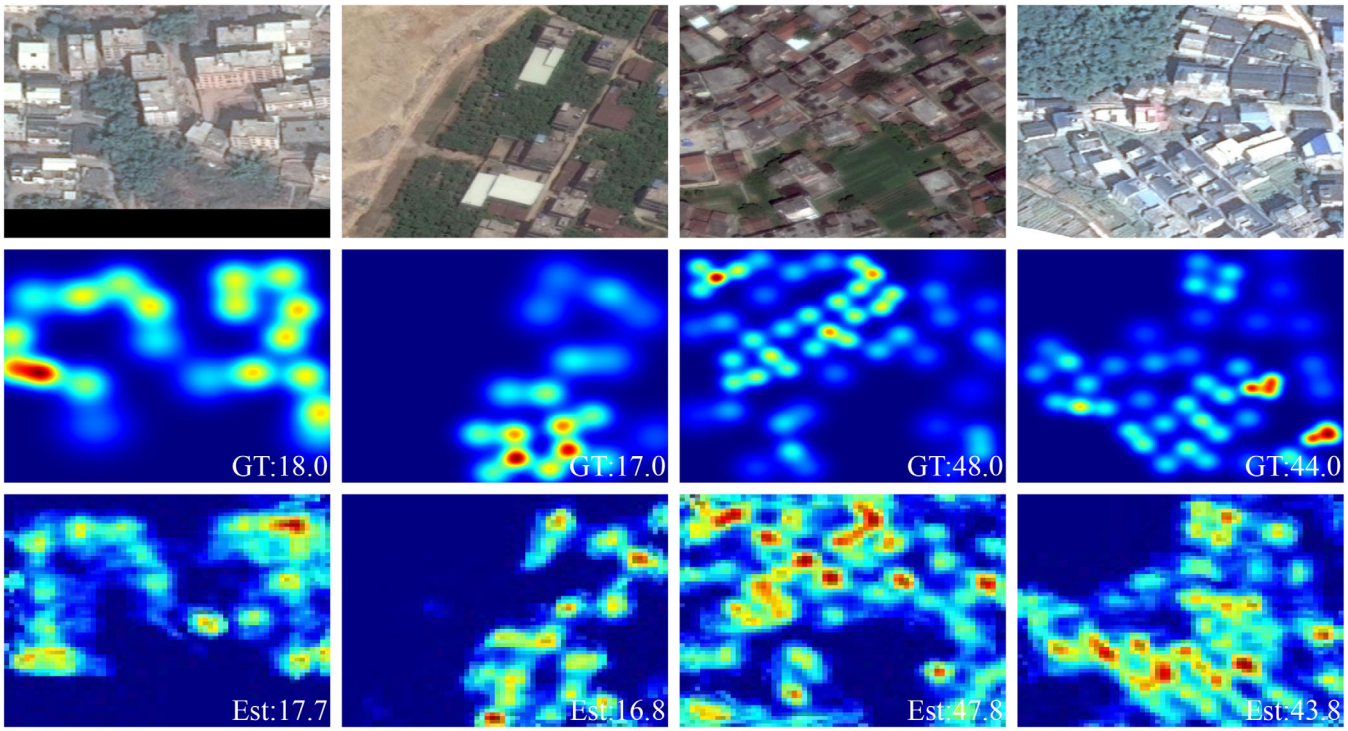

FIGURE 8 Subjective results on the RSOC_buiding dataset.

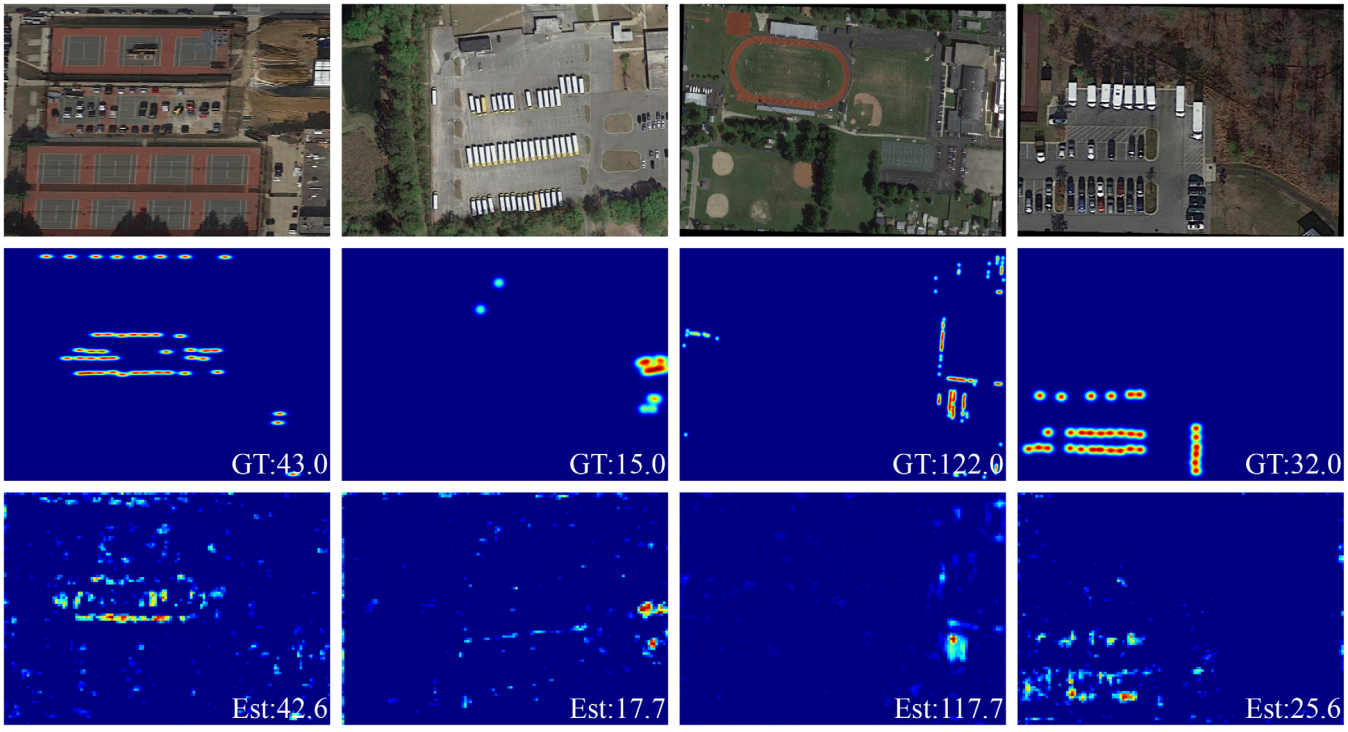

FIGURE 9 Subjective results on the RSOC_small-vehivle dataset.

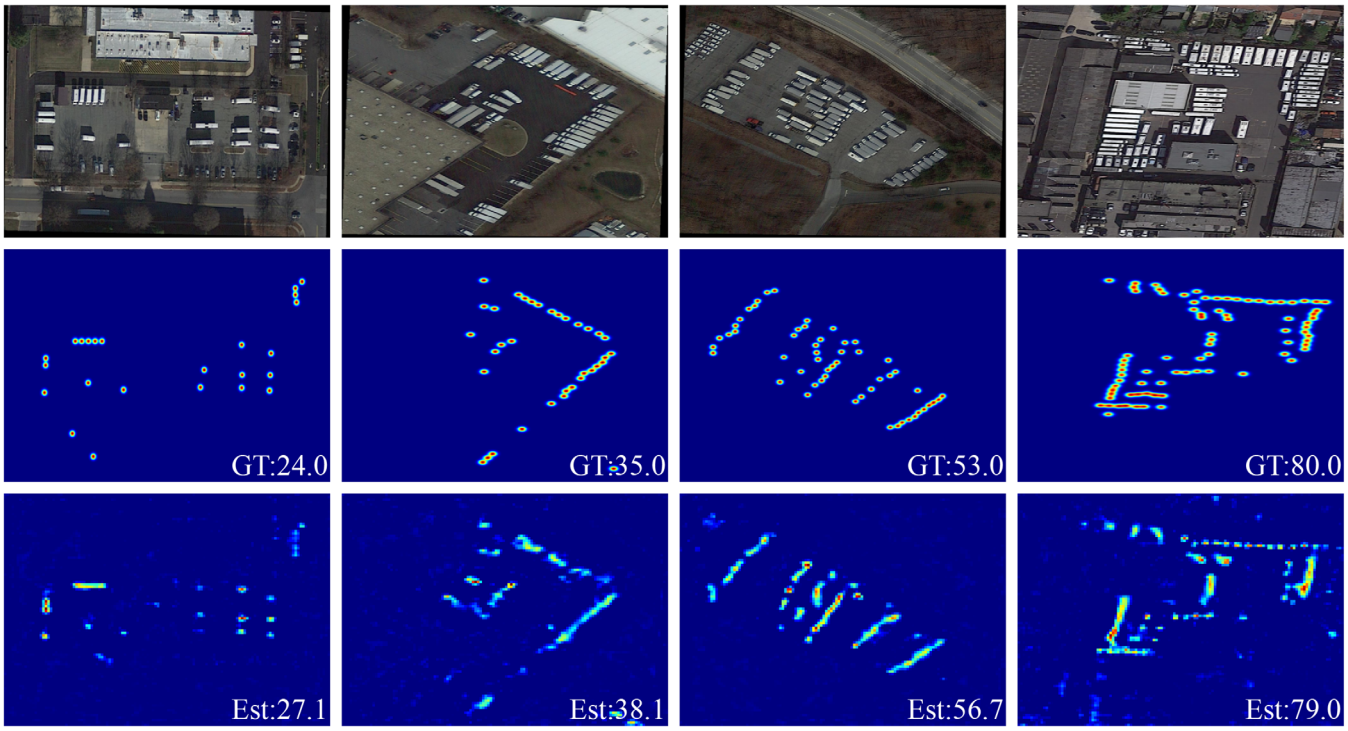

FIGURE 10 Subjective results on the RSOC_large-vehicle dataset.

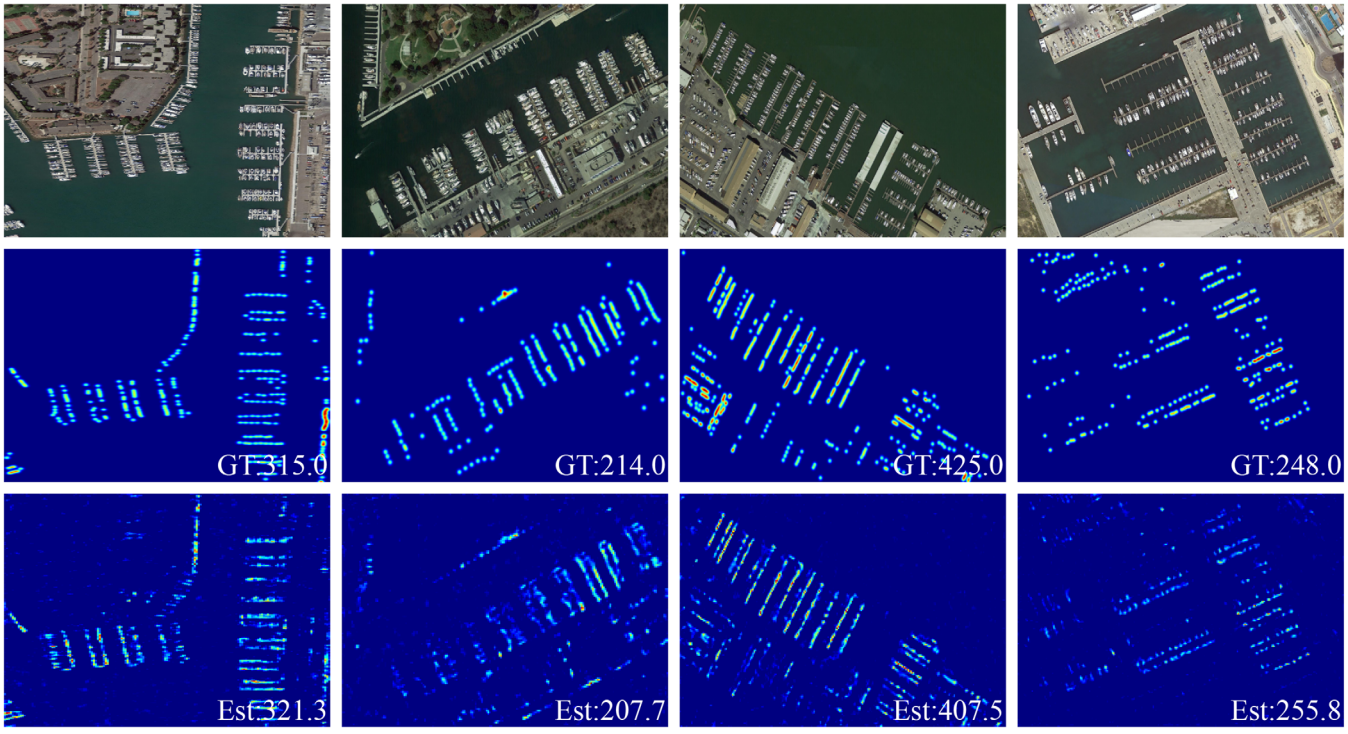

FIGURE 11 Subjective results on the RSOC_ship dataset.

4.5 | Experiments on CARPK and PUCPR+ datasets

Table 3 displays the objective comparison results on CARPK[46] and PUCPR+ datasets.[46] It proves that the SSFPNet outperforms other methods and achieves the highest performance on the CARPK[46] dataset. It shows a significant improvement of 11.73% in MAE and 19.45% in RMSE compared to TASNet.[28] Figure 12 depicts the subjective results obtained on the CARPK dataset. It is evident that the predicted density map accurately reflects the distribution of targets in the image and provides an accurate estimation of the target count. It is worth noting that, according to the previous work,[28,38,45,52] the predicted results on CARPK and PUCPR+ datasets are all integers. Therefore, we have unified the predicted vehicle counts as integers, and the results are shown in the third row of Figures 12 and 13.

From the performance results on PUCPR+ dataset[46] can be seen in Table 3, it is evident that SSFPNet has achieved a significant improvement in counting accuracy. Compared to TASNet,[28] SSFPNet has achieved an improvement of 44.19% in MAE and 39.13% in MSE. The visualizations in Figure 13 further demonstrate the stability and effectiveness of SSFPNet in various weather conditions. These visualizations further emphasize the consistent performance of SSFPNet and its ability to accurately count objects in challenging environmental conditions.

4.6 | Ablation studies

In this section, ablation studies are carried out to demonstrate the effectiveness of the PA module and HFP module in SSFNet.

Ablation study on the pyramid attention module: In order to verify the effect of different combinations of PSA and PCA on the model performance. A series of experiments are performed on the RSOC_ship dataset. The objective comparison results are shown in Table 4.

1. Baseline: The baseline model refers to the model consisting only of VGG-16 and the backend module, without the PA module and HFP module. The results obtained from this baseline model are the worst among all methods.
2. Baseline+PGC+PCA: The baseline model combined with the PA module, containing the PGC unit and the PCA unit.
3. Baseline+PGC+PSA: The baseline model combined with the PA module containing the PGC unit and the PSA unit.
4. Baseline+PGC+PSA||PCA: The baseline model with parallel connections of the PSA and PCA units in the PA module.

TABLE 3 Comparison results on the CARPK and PUCPR+ datasets.

	CARPK		PUCPR+	
Methods	**MAE**	**RMSE**	**MAE**	**RMSE**
YOLO12,[12]	102.89	110.02	156.72	200.54
FRCN53,[53]	103.48	110.64	156.76	200.59
LEP54,[54]	51.83	-	15.17	-
LPN46,[46]	23.80	36.79	22.76	34.46
SSD55,[55]	37.33	42.32	119.24	132.22
RetinaNet56,[56]	16.62	22.30	24.58	33.12
One-Look Regression[57]	59.46	66.84	21.88	36.73
MCNN[32]	39.10	43.30	21.86	29.53
CSRNet[34]	11.48	13.32	8.65	10.24
BL[58]	9.58	11.38	6.54	8.13
PSGCNet[59]	8.15	10.46	5.24	7.36
TASNet[28]	7.16	10.23	5.16	6.67
SSFPNet(Ours)	**6.32**	**8.24**	**2.88**	**4.06**

Note: The best results are highlighted in bold.

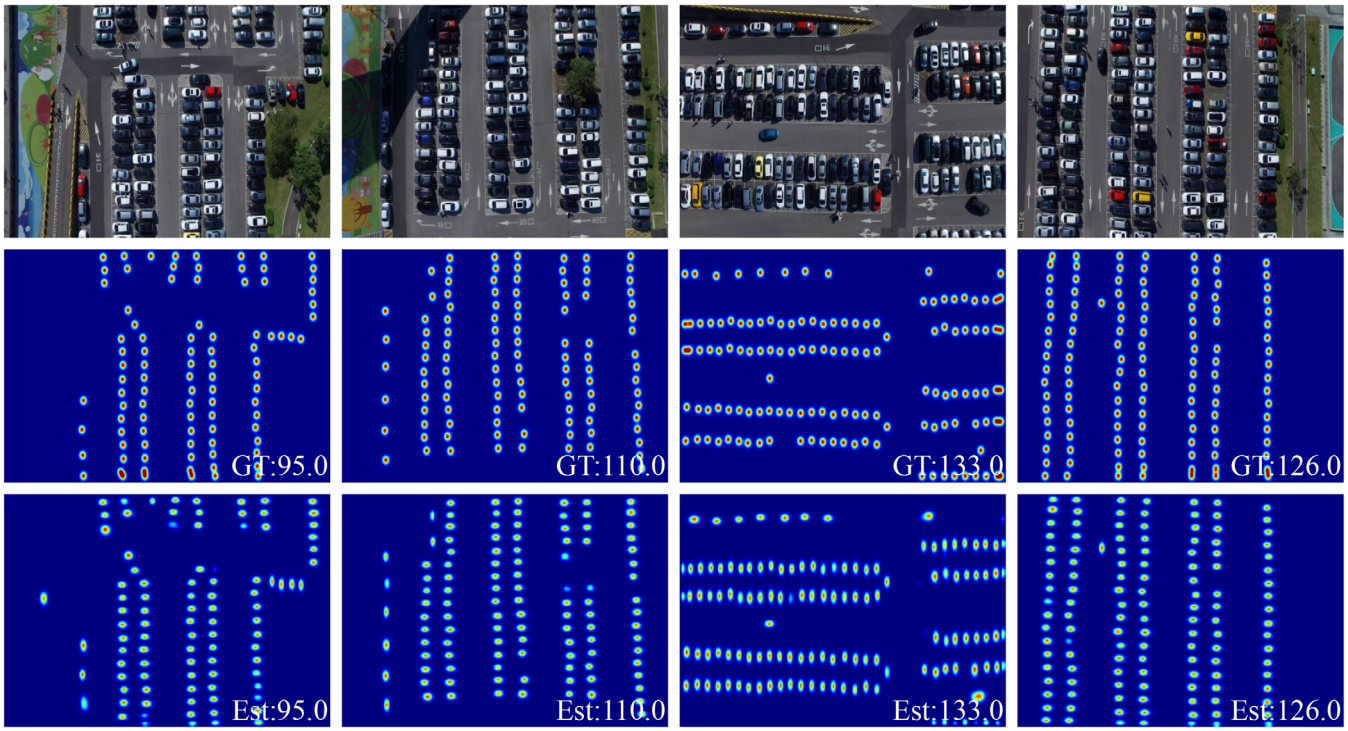

FIGURE 12 Subjective results on the CARPK dataset.

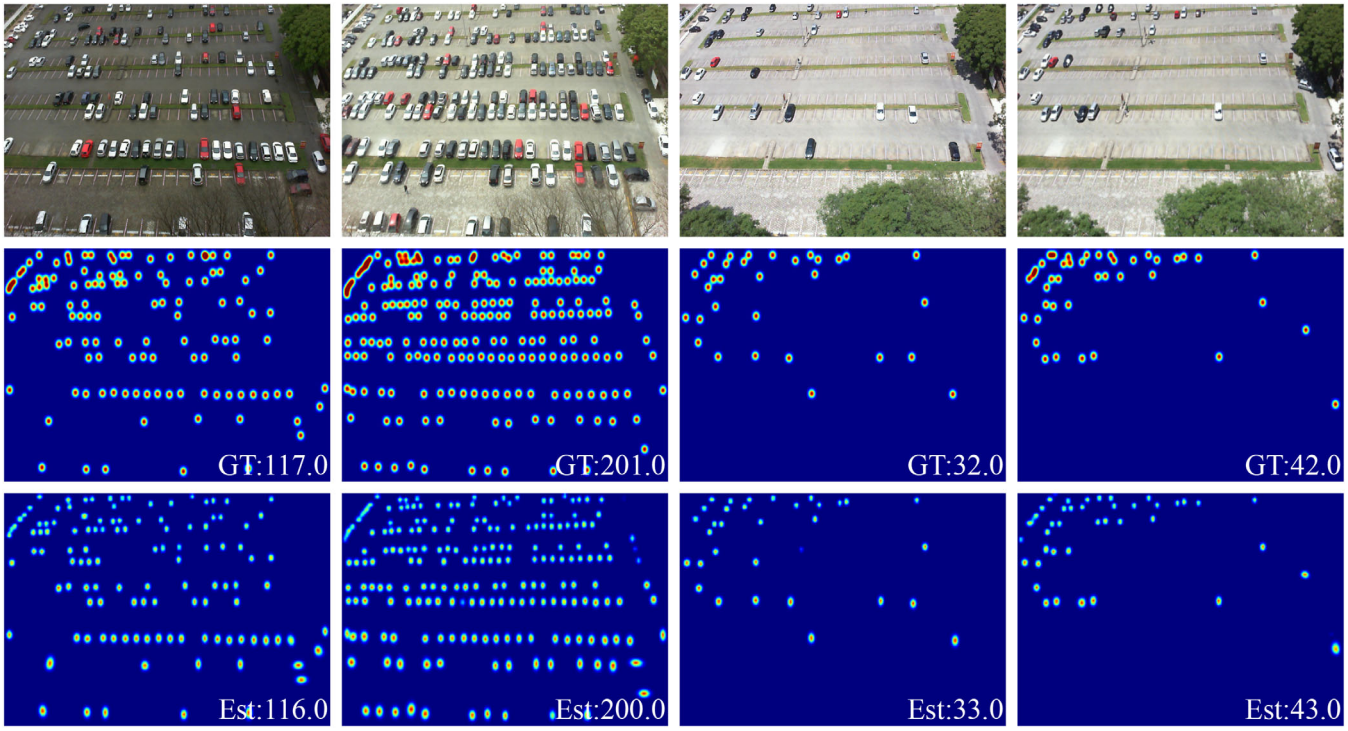

FIGURE 13 Subjective results on the PUCPR+ dataset.

TABLE 4 Comparative results of baseline with different PA combinations on the ship dataset.

Methods	MAE	RMSE
Baseline	181.44	221.89
Baseline+PGC+PCA	119.09	151.25
Baseline+PGC+PSA	151.09	186.55
Baseline+PGC+PSA||PCA	141.90	179.18
Baseline+PGC+PCA_PSA	145.21	181.47
Baseline+PGC+PSA_PCA	**87.67**	**117.82**

Note: The best results are emphasized with bold formatting.

TABLE 5 Objective comparison of different frequency domain branch numbers on the ship dataset.

Methods	MAE	RMSE
Baseline	181.44	221.89
Baseline+HFP(3,1)	170.90	230.13
Baseline+HFP(2,2)	125.53	168.10
Baseline+HFP(1,3)	**102.79**	**129.01**
Baseline+HFP(0,4)	117.76	162.65

Note: The best results are emphasized in bold font.

5. Baseline+PGC+PCA_PSA: The baseline model with sequential application of the PCA and PSA units in the PA module, in the order of PCA first and PSA second.
6. Baseline+PGC+PSA_PCA: The baseline model with sequential application of the PSA and PCA units in the PA module, in the order of PSA first and PCA second.

Among the different configurations, the Baseline+PGC+PSA_PCA stands out as the top performer, achieving the lowest MAE and RMSE values. This configuration incorporates the PSA unit followed by the PCA unit in the PA module. This specific order of applying attention mechanisms is crucial for optimizing counting accuracy. The initial PSA unit focuses on capturing spatial dependencies, while the subsequent PCA unit refines the feature representation by attending to channel-level information. This sequential arrangement allows for a hierarchical and progressive refinement of features, resulting in improved model performance. Thus, the strategic placement of PSA before PCA in the Baseline+PGC+PSA_PCA configuration plays a vital role in enhancing object counting capabilities.

Ablation study on frequency branches: The hybrid feature pyramid module consists of spatial and frequency domain pyramids. In order to verify the effectiveness of two pyramids with different domains. An ablation study is performed on the number of branched in the frequency domain. The objective comparison of different configurations is presented in Table 5.

As shown in Table 5, $HFP(i, j)$ indicates that in the HFP module, the number of spatial domain branches is i and the number of frequency domain branches is j. As the number of frequency domain branches increases, we observe an improvement in counting performance. However, if all branches are exclusively in the frequency domain, there is a slight decrease in MAE and MSE due to the lack of spatial scale information. Therefore, we adopt the optimal configuration of 1 spatial domain branch and 3 frequency domain branches, specifically Baseline+HFP(1,3). To fully leverage the hierarchical nature of the hybrid feature pyramid, the dilation rate of the first spatial domain branch is set to 2, while the frequency domain coefficients for the subsequent 3 branches are chosen as 16, 64, and 256, respectively. This choice retains the low-frequency scale information at 1/16, 1/64, and 1/256. It highlights the importance of preserving different frequency scales in the feature pyramid.

Ablation study on the pivotal components: Ablation studies are conducted on the ship dataset to investigate the individual effects and mutual influence of the two pivotal modules, that is, PA and HPF. The objective results with various configurations are presented in Table 6. It can be observed that the PA module significantly

TABLE 6 Objective results of the effects and interactions of the two pivotal modules on the ship dataset.

Methods	MAE	RMSE
Baseline	181.44	221.89
Baseline+PA	87.67	117.82
Baseline+HFP	102.79	129.01
Baseline+PA+HFP	**73.59**	**96.64**

Note: The best results are emphasized with bold formatting.

TABLE 7 Analysis of parameters (Params) and multiply-accumulates (MACs) in pivotal units of SSFPNet (with an input resolution of 512×512).

Methods	Params(M)	MACs(G)
Baseline	16.41	109.04
Baseline+PGC	18.14	116.14
Baseline+PGC+PSA_PCA	18.14	116.14
Baseline+HFP	19.82	123.01
Baseline+PGC+PSA_PCA+HFP	21.55	130.11

mitigates background interference on the ship dataset, demonstrating its powerful feature discrimination and representation capabilities. When combined with the HPF module, which further refines scale features, the integration of both modules harnesses their respective strengths and effectively addresses the impact of irrelevant backgrounds and scale variations. We provide an analysis of the parameters (Params) and multiply-accumulates (MACs) for the pivotal units of SSFPNet, and the results are shown in Table 7. It can be observed that the PSA and PCA units in the PA module do not increase the number of parameters and computational workload. Furthermore, it is worth noting that the introduction of the HFP module does not significantly increase the computational burden or model complexity.

5 | CONCLUSION AND FUTURE WORK

In this paper, we propose the SSFPNet for remote sensing object counting, which is used to weaken the influence of background and scale variation to improve the counting accuracy. The proposed SSFPNet network consisted of two primary modules, that is, a PA module and an HFP module. By simultaneously leaning spatial and channel attention weights on four branches, the PA module extracts rich features while highlighting significant features and reducing the sensitivity to the background. The HFP module combines and complements different scale information in the spatial domain and frequency domain, which improves the multiscale expression ability. The experimental results on benchmark remote sensing datasets RSOC, CARPK, and PUCPR+ confirm the effectiveness and superiority of SSFPNet. In the future, due to the presence of multiple object categories in remote sensing images, there is expected to be more attention and effort devoted to the research on handling multi-class object counting problems.

AUTHOR CONTRIBUTIONS

Conceptualization: Jinyong Chen, Mingliang Gao, and Gwanggil Jeon; *Methodology*: Jinyong Chen and Mingliang Gao; *Software*: Jinyong Chen and Wenzhe Zhai; *Validation*: Jinyong Chen and Mingliang Gao; *Formal Analysis*: Jinyong Chen and Mingliang Gao; *Investigation*: Jinyong Chen and Xiangyu Guo; *Resources*: Mingliang Gao and Qilei Li; *Writing*: Jinyong Chen; *Supervision*: Mingliang Gao, Gwanggil Jeon, and Qilei Li; *Project Administration*: Jinyong Chen, Gwanggil Jeon, and Mingliang Gao. All authors have reviewed and approved the published version of the manuscript.

FUNDING INFORMATION
There is no funding to support this manuscript.

CONFLICT OF INTEREST STATEMENT
The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT
Data sharing is not applicable to this article, as no datasets were generated or analyzed during the current study.

ORCID

Xiangyu Guo https://orcid.org/0000-0001-9405-3792

REFERENCES
1. Lian Z, Wang L. A novel forgery classification method based on multi-scale feature capsule network in mobile edge computing. *Softw Pract Exp*. 2024;54(9):1651-1670. doi:10.1002/spe.3245
2. Aghazadeh R, Shahidinejad A, Ghobaei-Arani M. Proactive content caching in edge computing environment: a review. *Softw Pract Exp*. 2023;53(3):811-855.
3. Candal-Ventureira D, González-Castaño FJ, Gil-Castiñeira F, Fondo-Ferreiro P. Is the edge really necessary for drone computing offloading? An experimental assessment in carrier-grade 5G operator networks. *Softw Pract Exp*. 2023;53(3):579-599.
4. Dinh DL, Nguyen HN, Thai HT, Le KH. Towards AI-based traffic counting system with edge computing. *J Adv Trans*. 2021;2021:1-15.
5. Mao Y, You C, Zhang J, Huang K, Letaief KB. A survey on mobile edge computing: the communication perspective. *IEEE Commun Surv Tutorials*. 2017;19(4):2322-2358.
6. Gao J, Gong M, Li X. Global multi-scale information fusion for multi-class object counting in remote sensing images. *Remote Sens*. 2022;14:4026.
7. Gao G, Liu Q, Wang Y. Counting dense objects in remote sensing images. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE; 2020:4137-4141.
8. Li W, Fu H, Yu L, Cracknell A. Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sens*. 2016;9(1):22.
9. Chen Z, Deng L, Luo Y, et al. Road extraction in remote sensing data: A survey. *Int J Appl Earth Observat Geoinformat*. 2022;112:102833.
10. Rathore MM, Ahmad A, Paul A, Rho S. Urban planning and building smart cities based on the internet of things using big data analytics. *Comput Netw*. 2016;101:63-80.
11. Gao G, Liu Q, Wang Y. Counting from sky: A large-scale data set for remote sensing object counting and a benchmark method. *IEEE Trans Geosci Remote Sens*. 2020;59(5):3642-3655.
12. Redmon J, Divvala SK, Girshick RB, Farhadi A. You only look once: unified, real-time object detection. *IEEE Conference Computer Vision Pattern Recognition (CVPR)*. IEEE; 2016:779-788.
13. Girshick RB. Fast R-CNN. *IEEE International Conference Computer Vision (ICCV)*. IEEE; 2015:1440-1448.
14. Pham VQ, Kozakaya T, Yamaguchi O, Okada R. COUNT Forest: CO-voting uncertain number of targets using random Forest for crowd density estimation. *IEEE International Conference Computer Vision (ICCV)*. IEEE; 2015:3253-3261.
15. Gao M, Souri A, Zaker M, Zhai W, Guo X, Li Q. A comprehensive analysis for crowd counting methodologies and algorithms in internet of things. *Clust Comput*. 2023. doi:10.1007/s10586-023-03987-y
16. Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans Pattern Anal Mach Intell*. 2011;34(4):743-761.
17. Paszke A, Gross S, Chintala S, et al. Automatic Differentiation in Pytorch. 2017.
18. Chan AB, Vasconcelos N. Bayesian poisson regression for crowd counting. *2009 IEEE 12th International Conference on Computer Vision(ICCV)*. IEEE; 2009:545-551.
19. Gao G, Gao J, Liu Q, Wang Q, Wang Y. CNN-based density estimation and crowd counting: a survey. *arXiv* 2020 abs/2003.12783.
20. Guo X, Gao M, Zhai W, Li Q, Kim KH, Jeon G. Dense attention fusion network for object counting in IoT system. *Mobile Networks Appl*. 2023;28:359-368.
21. Rong L, Li C. Coarse- and fine-grained attention network with background-aware loss for crowd density map estimation. *IEEE Winter Conference Applications of Computer Vision (WACV)*. IEEE; 2021:3674-3683.
22. Dijkstra K, van de Loosdrecht J, Atsma WA, Schomaker L, Wiering MA. CentroidNetV2: A hybrid deep neural network for small-object segmentation and counting. *Neurocomputing*. 2021;423:490-505.
23. Loh DR, Yong WX, Yapeter J, Subburaj K, Chandramohanadas R. A deep learning approach to the screening of malaria infection: automated and rapid cell counting, object detection and instance segmentation using mask R-CNN. *Comput Med Imaging Graph Official J Comput Med Imaging Soc*. 2021;88:101845.
24. Xu J, Le HM, Samaras D. Learning from pseudo-labeled segmentation for multi-class object counting. arXiv 2023;abs/2307.07677.
25. Guo X, Song K, Zhai W, Gao M, Li Q, Jeon G. Crowd counting in smart city via lightweight ghost attention pyramid network. *Futur Gener Comput Syst*. 2023;147:328-338.

26. Zhai W, Pan J, Li Q, Zou G, Yin L, Gao M. A channel-aware attention network for crowd counting. *2021 China Automation Congress (CAC)*. IEEE; 2021:4048-4052.
27. Cholakkal H, Sun G, Khan FS, Shao L. Object counting and instance segmentation with image-level supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (ICCV)*. IEEE; 2019:12397-12405.
28. Guo X, Anisetti M, Gao M, Jeon G. Object counting in remote sensing via triple attention and scale-aware network. *Remote Sens*. 2022;14(24):6363.
29. Guo X, Gao M, Zhai W, Li Q, Pan J, Zou G. Multiscale aggregation network via smooth inverse map for crowd counting. *Multimed Tools Appl*. 2022. doi:10.1007/s11042-022-13664-8
30. Zhai W, Gao M, Guo X, Li Q. Scale-context perceptive network for crowd counting and localization in Smart City system. *IEEE Internet Things J*. 2023;10(21):18930-18940.
31. Yu X, Liang Y, Lin X, Wan J, Wang T, Dai HN. Frequency feature pyramid network with global-local consistency loss for crowd-and-vehicle counting in congested scenes. *IEEE Trans Intell Transp Syst*. 2022;23(7):9654-9664.
32. Zhang Y, Zhou D, Chen S, Gao S, Ma Y. Single-image crowd counting via multi-column convolutional neural network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016:589-597.
33. Chen X, Bin Y, Sang N, Gao C. Scale pyramid network for crowd counting. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE; 2019:1941-1950.
34. Li Y, Zhang X, Chen D. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2018:1091-1100.
35. Gao J, Wang Q, Yuan Y. SCAR: spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing*. 2019;363:1-8.
36. Zhai W, Li Q, Zhou Y, et al. DA2Net: a dual attention-aware network for robust crowd counting. *Multimedia Systems*. 2022;29(5):3027-3040.
37. Jiang X, Zhang L, Xu M, et al. Attention scaling for crowd counting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*; 2020:4705-4714.
38. Zhai W, Gao M, Souri A, et al. An attentive hierarchy ConvNet for crowd counting in smart city. *Clust Comput*. 2023;26(2):1099-1111.
39. Idrees H, Saleemi I, Seibert C, Shah M. Multi-source multi-scale counting in extremely dense crowd images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2013:2547-2554.
40. Ehrlich M, Davis LS. Deep residual learning in the jpeg transform domain. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2019:3484-3493.
41. Xu K, Qin M, Sun F, Wang Y, Chen YK, Ren F. Learning in the frequency domain. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2020:1740-1749.
42. Qin Z, Zhang P, Wu F, Li X. Fcanet: frequency channel attention networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2021:783-792.
43. Shu W, Wan J, Tan KC, Kwong S, Chan AB. Crowd counting in the frequency domain. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2022:19618-19627.
44. Guo X, Gao M, Zhai W, Shang J, Li Q. Spatial-frequency attention network for crowd counting. *Big Data*. 2022; 10(5):453-465.
45. Zhai W, Gao M, Li Q, Jeon G, Anisetti M. FPANet: feature pyramid attention network for crowd counting. *Appl Intell*. 2023;53:19199-19216.
46. Hsieh MR, Lin YL, Hsu WH. Drone-based object counting by spatially regularized regional proposal network. *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE; 2017:4165-4173.
47. Sindagi V, Patel V. CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE; 2017:1-6.
48. Cao X, Wang Z, Zhao Y, Su F. Scale aggregation network for accurate and efficient crowd counting. *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer; 2018.
49. Wang Q, Gao J, Lin W, Yuan Y. Learning from synthetic data for crowd counting in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2019:8190-8199.
50. Liu W, Salzmann M, Fua P. Context-aware crowd counting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2019:5094-5103.
51. Zhu L, Zhao Z, Lu C, Lin Y, Peng Y, Yao T. Dual path multi-scale fusion networks with attention for crowd counting. ArXiv. 2019; abs/1902.01115.
52. Kilic E, Ozturk S. An accurate car counting in aerial images based on convolutional neural networks. *J Ambient Intell Human Comput*. 2021;14:1259-1268.
53. Ren S, He K, Girshick RB, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Tpami*. 2015;39:1137-1149.
54. Stahl T, Pintea SL, Gemert JCV. Divide and Count: generic object counting by image divisions. *IEEE Trans Image Process*. 2019;28:1035-1044.
55. Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector. *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer; 2016:21-37.
56. Lin TY, Goyal P, Girshick RB, He K, Dollár P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. 2020;42:318-327.
57. Mundhenk TN, Konjevod G, Sakla WA, Boakye K. A large contextual dataset for classification, detection and counting of cars with deep learning. *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer; 2016.

58. Ma Z, Wei X, Hong X, Gong Y. Bayesian loss for crowd Count estimation with point supervision. *Proceedings of the IEEE/CVF International Conference Computer Vision (ICCV)*. IEEE; 2019:6141-6150.
59. Gao G, Liu Q, Hu Z, Li L, Wen Q, Wang Y. PSGCNet: a pyramidal scale and global context guided network for dense object counting in remote-sensing images. *IEEE Trans Geosci Remote Sens*. 2022;60:1-12.

How to cite this article: Chen J, Gao M, Guo X, Zhai W, Li Q, Jeon G. Object counting in remote sensing via selective spatial-frequency pyramid network. *Softw: Pract Exper*. 2024;54(9):1754-1773. doi: 10.1002/spe.3287