Mobile Networks and Applications (2023) 28:359–368
https://doi.org/10.1007/s11036-023-02090-1

Dense Attention Fusion Network for Object Counting in IoT System

Xiangyu Guo[1] · Mingliang Gao[1] · Wenzhe Zhai[1] · Qilei Li[2] · Kyu Hyung Kim[3] · Gwanggil Jeon[4]

Accepted: 14 October 2022 / Published online: 25 January 2023

Abstract
IoT has been overwhelmingly empowered by the rapid development of big-data ecosystems, such as remote sensing technology which runs all the time in obtaining accurate and high-quality images to facilitate the subsequent image processing and content analysis in embedded devices. Object counting, which aims to estimate the number of objects in a captured image, is one of the most crucial tasks among multimedia data and wireless network. However, there are enormous inherent factors that seriously degrade the counting performance in remote sensing, e.g. the background clutter, scale variation, and orientation arbitrariness. In this paper, we tackle the aforementioned problems in a divide-and-conquer manner by devising the dense attention fusion network (DAFNet). Specifically, we introduce an iterative attention fusion (IAF) module, which mainly relies on the multiscale channel attention (MCA) unit, to alleviate the side effect caused by background clutter. Meanwhile, to overcome the intrinsic scale variations, we build a dense spatial pyramid (DSP) module to consider the hierarchical information obtained under diverse receptive fields. Finally, we stack deformable convolution layers to deal with the orientation arbitrariness. The synergy of the proposed IAF and DSP modules substantially promotes the effectiveness of the proposed DAFNet, which can be demonstrated by the notable superiority in extensive experiments on the remote sensing counting datasets against state-of-the-art competitors.

Keywords Remote sensing · IoT · Object counting · Embedded devices

1 Introduction

Internet of Things (IoT) network connects us and the world from every aspect. It significantly relies on large-scale data acquired from reliable sources, one of which is the remote sensing systems [1], owning to its numerous advantages, e.g. rich ground information and clear content, which can substantially facilitate the subsequent processing in the embedded devices. In recent years, widespread attention has been focused on object detection [2, 3], image segmentation [4, 5] in remote sensing images. However, object counting is still in its infancy. Counting the major objects on the ground (e.g. buildings, vehicles, and ships) in remote sensing images is instrumental for numerous research fields, e.g. urban planning [6], environment management [7] and digital city model construction [8].

The purpose of object counting is to estimate the object number in a given image. The research has yielded satisfactory results in many domains, i.e., crowd counting [9, 10], animal counting [11], cell counting [12], and vehicle counting [13]. The pioneering object counting methods are mainly detection-based [14, 15] methods, which utilize a detector to identify and locate the position of an object. The limitation of these approaches is that they only handle the sparse scene and performs barely satisfactory in dense condition. To overcome these disadvantages, regression-based methods [16, 17] have been proposed. The key idea of these methods is learning a mapping from an image to the count number. Nevertheless, these methods tend to ignore the location information, thus are suboptimal for generating a density map.

Nowadays, the powerful ability of convolutional neural network (CNN) drives an increasing number of researchers

✉ Mingliang Gao
mlgao@sdut.edu.cn

1 School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China

2 School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK

3 Korea Conformity Laboratories, 199, Gasan digital 1-ro, Geumcheon-gu, Seoul, Republic of Korea

4 Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, Republic of Korea

to leverage it to improve counting accuracy. These CNN-based methods aim to regress a density map and sum up the all pixels to estimate the final counts. However, there still exists some challenges in this domain, *e.g.,* background clutter, scale variation and orientation arbitrariness. The remote sensing images are often captured from the sky, thus the objects present a tiny size. The complex background makes the model unable to emphasize the region of interest, which causes the objects to drown in background clutter. Furthermore, objects in remote sensing images show diverse sizes, ranging from several pixels to hundreds of pixels. The large-scale variation is unfavourable to predict an accurate number of objects. Besides, the objects in the remote sensing images possess uncertain orientation, which makes the remote sensing counting task more difficult than other counting tasks.

The challenges in remote sensing images are depicted in Fig. 1 (Note: notable regions are shown in red bounding boxes). Figure 1(a) exhibits the different scales of ships. Figure 1(b) shows that the vehicles are obscured by trees and suffered from background clutter. Figure 1(c) depicts the arbitrary orientations of vehicles and buildings.

To cope with these challenges and enhance the counting performance in remote sensing, we propose the dense attention fusion network (DAFNet). The VGG-16 is first utilized as the feature extractor to extract image features. Meanwhile, we propose an iterative attention fusion (IAF) module, which mainly based on multiscale channel attention (MCA) unit, to suppress the influence of the background clutter. In addition, we introduce a dense spatial pyramid (DSP) module to address the problem of scale variation. Finally, the deformable convolution layers are adopted to resolve the problem of orientation arbitrariness. In a nutshell, the main contributions are three aspects.

1. We develop the DAFNet, which is capable of boosting the counting performance of remote sensing objects in a divide-and-conquer manner.
2. We design an IAF module to suppress the influence of the background clutter, and build a DSP module to address the problem of scale variation.
3. We execute extensive experiments to verify the performance of object counting in challenging remote sensing scenarios. Meanwhile, detailed ablation studies are conducted to prove the effectiveness of the individual components in the proposed model.

2 Related work

2.1 Traditional object counting methods

The traditional object counting methods fall into two broad categories, i.e., detection-based [14, 15] and regression-based methods [16, 17]. The detection-based methods utilize a head detector to realize the counting task. Li et al. [18] utilized a histogram of oriented gradient (HOG) based head-shoulder detector to detect the number of individuals present in the given image. Ge et al. [19] deployed a marked point process based on Bayesian formulation to locate persons in congested scenes. The detection-based methods present satisfactory results in sparse crowds, but they have poor performance in dense scenes.

To deal with the problems in congested scenes, many regression-based methods have been proposed to enhance the counting accuracy, which mainly learn a mapping from the given image to the count. Lempitsky et al. [16] built a linear mapping between subtle features and the density map, which is obtained by understanding the spatial distribution.

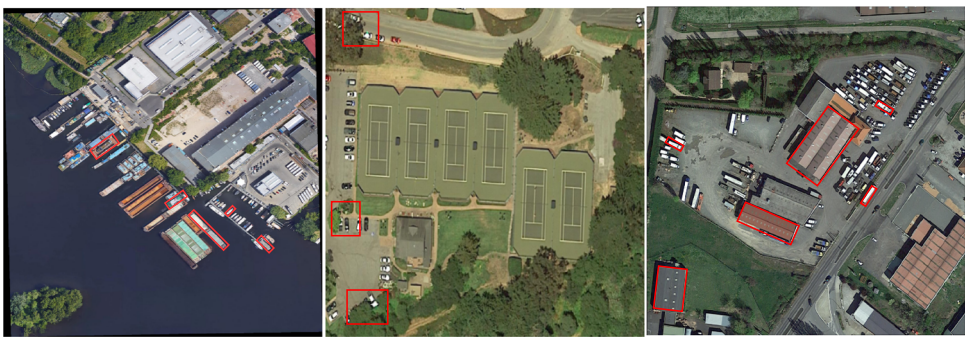

Fig. 1 Illustrations of the challenges in remote sensing images

Pham et al. [17] introduced a non-linear mapping, which utilizes a crowdedness prior to optimize two random forests. Although these methods have achieved gratifying results, their performance is greatly degraded by challenge factors including scale variation and background clutter.

2.2 CNN-based object counting methods

Benefitting by the advantages of deep learning, the CNN-based methods [20] have been widely used to improve counting accuracy in recent years. Zhang et al. [21] built a three-branch network named MCNN for crowd counting. Each branch in MCNN adopts different kernel sizes to obtain diverse receptive fields to cope with scale variations. Similar to MCNN, Li et al. [22] also proposed a multi-column network called congested scene recognition network (CSRNet), in which four parallel dilated convolutional layers are added to the back-end of the network to enlarge the receptive field. Albeit alleviating the influence of scale variation, the multi-column network is difficult to train since it requires more time and more bloated structure. Besides, it can lead to information redundancy as the multi-column network using different branches but almost the same network structures [20, 22].

To avoid the above problems, many advanced approaches have exploited multiscale feature fusion to alleviate the side effect caused by scale variation. Liu et al. [23] built a content-aware network (CANet), which employs different receptive fields to fuse features. Gao et al. [24] put forward a perspective crowd counting net (PCCNet) aiming to tackle the perspective variations. The network can merge hierarchical features with the assistance of density map estimation module and foreground segmentation module.

In recent years, attention mechanism has been adopted to counting task, and achieved extraordinary performance [25–27]. Attention mechanism aims to deliberately emphasize the region of interest and minimize the impact of background noise. Liu et al. [28] introduced an attention map to compute the different crowd density levels and suppress the background noise. Sindagi et al. [29] employed a spatial attention module to choose the discriminative areas in feature maps, and a global attention module to suppress the unnecessary information along the channel dimension. Jiang et al.[30] introduced an attention scaling network to produce the scaling factors, which are capable to reflect the density levels. The scaling factors are multiplied by the mask density map generated by the front-end network to obtain a refined prediction. Rong et al. [31] built a crowd region recognizer (CRR), which can generate a coarse-grained attention map, to alleviate the background noise.

3 The proposed method

3.1 Overview

Figure 2 exhibits the framework of the DAFNet. It includes four modules: (1) A feature extractor module to obtain the basic features; (2) An IAF module to address the background clutter; (3) A DSP module to deal with the scale

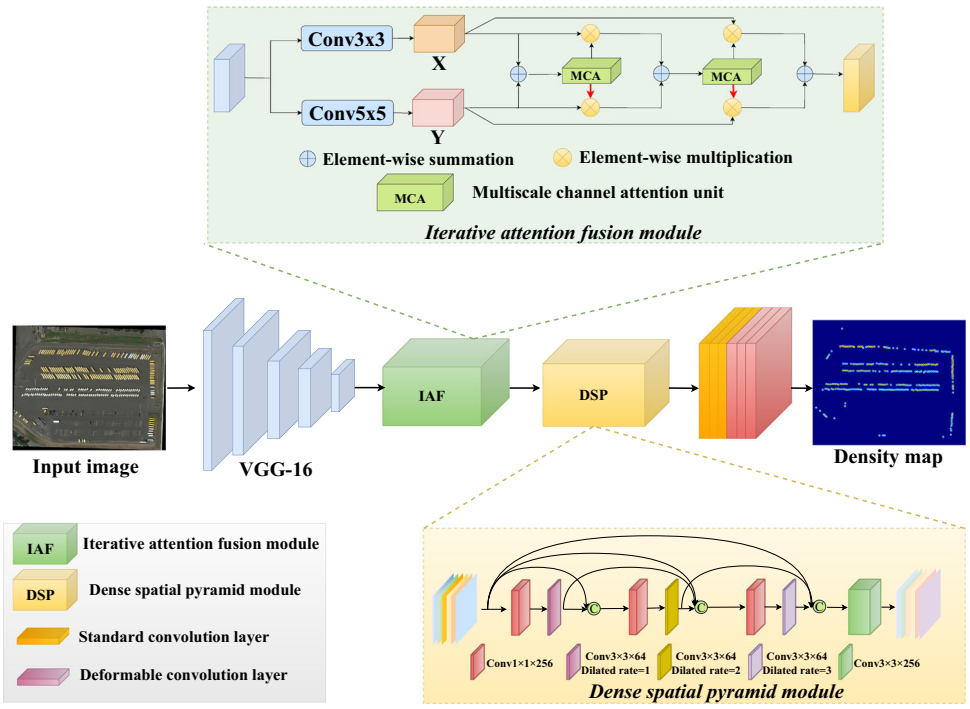

Fig. 2 Framework of the DAFNet

variation; (4) A deformable convolution module to solve the orientation arbitrariness.

The first ten convolution layers of VGG-16 are utilized to extract the low-level features. The features are fed into the IAF module to suppress the effect of background and then produce the refined feature map. Meanwhile, the DSP module is introduced to perceive the scale through the various dilated convolution layers. Finally, the deformable convolution layers are employed to overcome the orientation arbitrariness. The DAFNet can be represented as follows,

$$M = Dconv(f_{\text{dsp}}(f_{\text{iaf}}(\mathbf{X}))), \tag{1}$$

where M and $\mathbf{X}$ denote the prediction and the low-level features, respectively. The f_{dsp} and f_{iaf} are the functions of the DSP and IAF modules. $Dconv$ represents the deformable convolution layers.

3.2 Iterative attention fusion module

The IAF module is built to suppress the background clutter. It is achieved by two multiscale channel attention (MCA) units. The framework of the MCA unit is illustrated in Fig. 3. The top-branch adopts two 1×1 convolution layers to enrich the input feature $X \in \mathbb{R}^{C\times H\times W}$ and outputs a local attention feature map $M_l \in \mathbb{R}^{C\times H\times W}$. The operation of the top branch is formulated as,

$$M_l = \text{BN}(\text{Conv}_2(\text{ReLU}(\text{BN}(\text{Conv}_1(\mathbf{X}))))), \tag{2}$$

where BN is short for batch normalization and ReLU denotes the activation function. Compared with the top branch, the mid-branch aims to achieve the global context aggregation and obtain a global attention feature map $M_g \in \mathbb{R}^{C\times1\times1}$, which is helpful to highlight the global context information. The M_g is computed as,

$$M_g = \text{GAP}(\text{BN}(\text{Conv}_2(\text{ReLU}(\text{BN}(\text{Conv}_1(\mathbf{X})))))), \tag{3}$$

where GAP represents a global average pooling operation.

It is remarkable that the local attentional map M_l has the different dimensions with the global attentional map M_g. This is due to that the GAP operation compresses the input feature map to 1×1 in spatial dimension. Subsequently, the local and global feature maps are aggregated by the sum operation to assemble the inner discriminative information. Then, the fused feature is activated by the Sigmoid function in spatial dimension while the channel dimension remains unchanged.

The purpose of the bottom branch is to aggregate the original features to produce a high-quality density map that contains rich local and global context information. In a nutshell, the refined feature map M' by the MCA unit is defined as

$$M' = X \otimes \text{Sigmoid}(L(M) \oplus G(M)), \tag{4}$$

where M and M' denote the input and refined feature maps, respectively. $L(\cdot)$ and $G(\cdot)$ represent the local and global attention functions. $\otimes$ denotes the element-wise multiplication, and $\oplus$ is the sum operation. By enhancing the features through MCA unit, the object area can be better distinguished.

Owing to the merits of the MCA unit, we further assemble it to build the IAF module. As depicted in Fig. 2, the IAF module aims to fuse two input feature maps, i.e., X and Y, which are obtained by convolution layers with diverse kernel sizes. Such strategy is different from the traditional fusion ways[21, 27], most of which only take one feature as the input for fusion process and adopt multiply branches to acquire different level features. First, X and Y are integrated into a feature map through the element-wise summation. The fused feature map is fed into the MCA unit to output two feature maps: One is obtained by multiplying the initial feature X with the weight σ, and the other is acquired by multiplying the initial feature Y with the weight $(1-\sigma)$ (denoted in red line in Fig. 2).

To fully exploit the context information, selecting a high-quality input is helpful to generate useful fusion weights. Hence, an intuitive way is to adopt another attention unit to fuse input features. It is an iterative process, so we name the module as iterative attention fusion (IAF) module. It is formulated as,

$$Z = M(X \oplus Y) \otimes X + (1 - M(X \oplus Y)) \otimes Y, \tag{5}$$

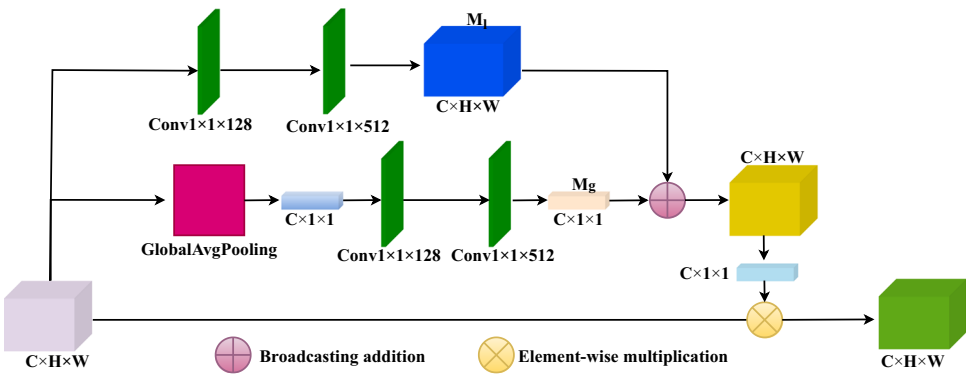

Fig. 3 Framework of the MCA unit

where Z represents the fused feature map. $X \oplus Y$ denotes the attentional feature fusion process of the first stage, which can be represented as,

$$X \oplus Y = M(X + Y) \otimes X + (1 - M(X + Y)) \otimes Y, \quad (6)$$

where $\oplus$ and $\otimes$ denote the element-wise summation and multiplication, respectively. $M(\cdot)$ refers to the function of the MCA unit.

3.3 Dense spatial pyramid module

In remote sensing images, the object often suffered from scale variation in dense regions. To cope with the problem, we introduce the dense spatial pyramid (DSP) module. The framework of DSP module is shown in Fig.4.

The DSP module mainly consists of standard and dilated convolution layers which are dense connected. The dilated convolution layers with different dilated rates aim to enlarge the receptive fields while saving the parameters. A convolution layer with kernel size 1×1 is added in front of each dilated convolution layer to reduce the number of channels. The 3×3 convolution layer is employed to integrate the previously generated features. In addition, the dense connection increases the number of channels by a factor of four, the convolution layer can also be used for channel reduction. Once these structures have been arranged, another significant issue is choosing suitable dilated rates for the dilated convolution layers.

Considering the objects in remote sensing images often appear tiny sizes, choosing large dilated rates may lead to the semantic information loss. This is adverse to capture the detailed features. In addition, since the change of object size in the image is continuous, it requires a dense sampling range. To this aim, the dilated rates are set as 1, 2 and 3, respectively. The parameter configuration can retain the spatial information of objects as much as possible.

As shown in Fig. 4, the DSP module can cover all pixel information of the input feature map. Meanwhile, as the chosen dilated rates are small, it can prevent the module from capturing the irrelevant information. Overall,

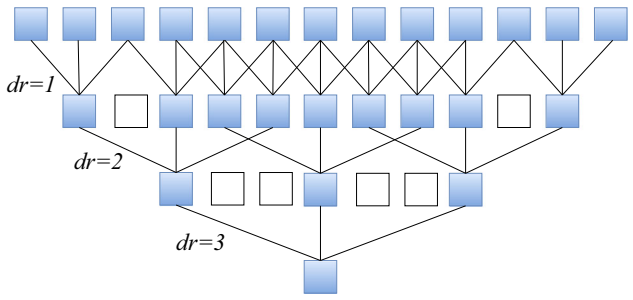

Fig. 4 Flowchart of DSP module for capturing information, where dr denotes the dilated rate. The black block denotes the source of information

the process of the DSP module is formulated as,

$$D = \mathrm{Conv}^{i}_{3\times3}(\mathrm{Conv}_{1\times1}(X))), i \in \{1, 2, 3\}, \quad (7)$$

$$O = \mathrm{Conv}_{3\times3}(\mathrm{DC}(D)), \quad (8)$$

where D and X represent the features after the dilated convolution layers and the input features, respectively. O refers to the output of the DSP module. DC denotes the dense connection, and i is the dilated rate.

The dilated convolution layers in the DSP module are densely connected with others. In this case, each layer can acquire the information from the previous layers, and transfer the information to the subsequent layers. This DSP module can increase the scale diversity to deal with scale variation.

3.4 Loss function

We utilize l_2 distance as the loss function to measure the discrepancy between the estimated and the ground truth density map. It is denoted as,

$$Loss = \frac{1}{\mathrm{N}} \left\| y - \hat{y} \right\|_2^2, \quad (9)$$

where N denotes the number of test images. y and $\hat{y}$ denote the estimated and the ground truth values, respectively.

3.5 Ground truth generation

The ground truth map M_{gt} is generated by adopting a Gaussian kernel convolving a delta function [21]:

$$M_{gt} = \sum_{i=1}^{H} \delta(x - x_i) * G_{\sigma_i}(x), \sigma_i = \beta \bar{d}_i, \quad (10)$$

where H denotes the number of head annotations and x refers to the position pixel. σ_i represents the variance of the Gaussian kernel. The $\delta(x - x_i)$ depicts a target head.

4 Experiments and analysis

4.1 Implementation details

The experiments are performed in the PyTorch framework [32]. We employ data augmentation to prevent overfitting. Specifically, for the building subdataset, the images are cropped to 256×256. For the other three subdatasets, e.g. small-vehicle, large-vehicle and ship subdatasets, the images are resized to 128×128. Meanwhile, we adopt a horizontal flip to double the data volumes.

During the training phase, we adopt the stochastic gradient descent (SGD) optimizer to train the proposed network in the end-to-end manner. We set the learning rate

as 1e-7, and the decay rate is set as 5e-4. The momentum of the SGD optimizer is set as 0.95. The maximum number of training epochs is set as 450. For building subdataset, the batch size is set to 16. Considering that the other three subdatasets have large resolutions, which may result in out-of-memory of GPU, the batch size is set to 4.

4.2 Evaluation metrics

The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are adopted for objective evaluation, which are formulated as,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|GT^i - Est^i|, \tag{11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|GT^i - Est^i|^2}, \tag{12}$$

where N denotes the number of testing samples. GT^i and Est^i represent the ground-truth and estimated count of the i-th sample, respectively. The two metrics can reflect the accuracy and robustness, respectively. Specifically, lower MAE and RMSE represent better counting accuracy and robustness of the model.

4.3 Comparative analysis

Experiments are conducted on RSOC dataset [32], which contains the most remote sensing images till now. According to different kinds of objects, the RSOC dataset is divided into four subdatasets, i.e., building, small-vehicle, large-vehicle, and ship. All the images are captured by satellites, and some samples are illustrated in Fig. 5. Comparative experiments are carried out among various state-of-the-art methods. The experimental results are reported in Table 1.

The building subdataset contains 2468 images crawling from Google Earth. The training and test sets have 1205 and 1263 images, respectively. The average resolution is 512 × 512. There are totally 76,215 annotated buildings in this dataset. For the building subdataset, the proposed DAFNet scores 7.39 and 10.87 in MAE and RMSE, respectively, both outperforming the competitors. Specifically, the second-best method SPN [10] built a scale pyramid to deal with scale variation by connecting parallel convolution layers, the proposed DAFNet address the problem by dense connection. Compared with it, the DAFNet decreases the MAE and RMSE by 4.5% and 5.3%, respectively.

The small-vehicle subdataset contains 280 images with 148,838 annotations in total (222 images for training and 58 images used for test). The average resolution of the images is 2473 × 2339. The subdataset is collected from the DOTA dataset. On this subdataset, the DAFNet gets a score of 404.46 and 1211.78 in MAE and the RMSE, both ranking the first place. Compared with the second-best method, i.e., SFANet [36], it decreases by 7.1% and 6.2% in terms of MAE and RMSE. Although the proposed method obtains the best results, the scores of MAE and RMSE are still very high, indicating that there is much room for improvement.

The large-vehicle subdataset is composed of 172 images with 16,594 annotated instances (108 images for training and 64 images used for test.). The average size of each image is 1552 × 1573. On the large-vehicle subdataset, the DAFNet achieve the best scores of 27.32 in MAE and 40.85 in RMSE. Compared with the SCAR [9], which also adopts the attention mechanism, the proposed DAFNet improves the MAE and RMSE by 56.5% and 48.7%.

The ship subdataset includes 137 images, of which 97 images are used for training and 40 images for test. Compared with other subdatasets, the images in this subdataset have the highest resolution 2558 × 2668. On this subdataset, the proposed DAFNet scores 200.43 and 294.11 in MAE and RMSE, both ranking the first place among the methods. Compared with the SPN [10], which is also built to overcome the scale variation in dense regions, the proposed DAFNet improves the MAE and RMSE by 16.9% and 25.1%, respectively.

Figure 6 provides some visualized samples. It can be seen that both the estimated density map and the estimated counting results closely approach to the ground truth.

4.4 Ablation study

To validate the effectiveness of proposed IAF and DSP modules in the DAFNet, we carry out ablation studies on the

Fig. 5 Samples in RSOC dataset. (a)∼(d) denote the images in building, small-vehicle, large-vehicle and ship subdatasets, respecitively

Table 1 Comparative results on the building, small-vehicle, large-vehicle, and ship datasets. The best performances are highlighted in **bold**

Method	Building		Small-vehicle		Large-vehicle		Ship	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN[21]	13.65	16.56	488.65	1317.44	36.56	55.55	263.91	412.30
CMTL[33]	12.78	15.99	490.53	1321.11	61.02	78.25	251.17	403.07
SANet[34]	29.01	32.96	497.22	1276.66	62.78	79.65	302.37	436.91
SCAR [9]	26.90	31.35	497.22	1276.65	62.78	79.64	302.37	436.92
CSRNet[22]	8.00	11.78	443.72	1252.22	34.10	46.42	240.01	394.81
SPN [10]	7.74	11.48	455.16	1252.92	36.21	50.65	241.43	392.88
CAN [23]	9.12	13.38	457.36	1260.39	34.56	49.63	282.69	423.44
SFCN [35]	8.94	12.87	440.70	1248.27	33.93	49.74	240.16	394.81
SFANet [36]	8.18	11.75	435.29	1284.15	29.04	47.01	201.61	332.87
DAFNet (Ours)	**7.39**	**10.87**	**404.46**	**1211.78**	**27.32**	**40.85**	**200.43**	**294.11**

small-vehicle subdataset. As shown in Table 1, the small-vehicle subdataset is the most challenging, as the MAE and RMSE scores in the small-vehicle subdataset are the highest. Therefore, the ablation experiments on this dataset is more objective. The detailed configurations are depicted as follows:

Baseline: The first ten layers of VGG-16 and the deformable convolution layers are employed.

Baseline+IAF: Adding the IAF module between the VGG-16 and the deformable layers.

Baseline+DSP: Adding the DSP module between the VGG-16 and the deformable layers.

Baseline+DSP+IAF: Adding the DSP and IAF modules sequentially between the VGG-16 and the deformable layers.

Baseline+IAF+DSP: Adding the IAF and DPS modules sequentially between the VGG-16 and the deformable layers.

The ablation experimental results are tabulated in Table 2. It can be observed that the baseline scores 498.85 and 1322.17 in MAE and RMSE, respectively. Employing the IAF module on the baseline basic, one can see that the MAE and RMSE decrease by 14.9% and 3.6%. This demonstrates the effectiveness of the IAF module. Besides,

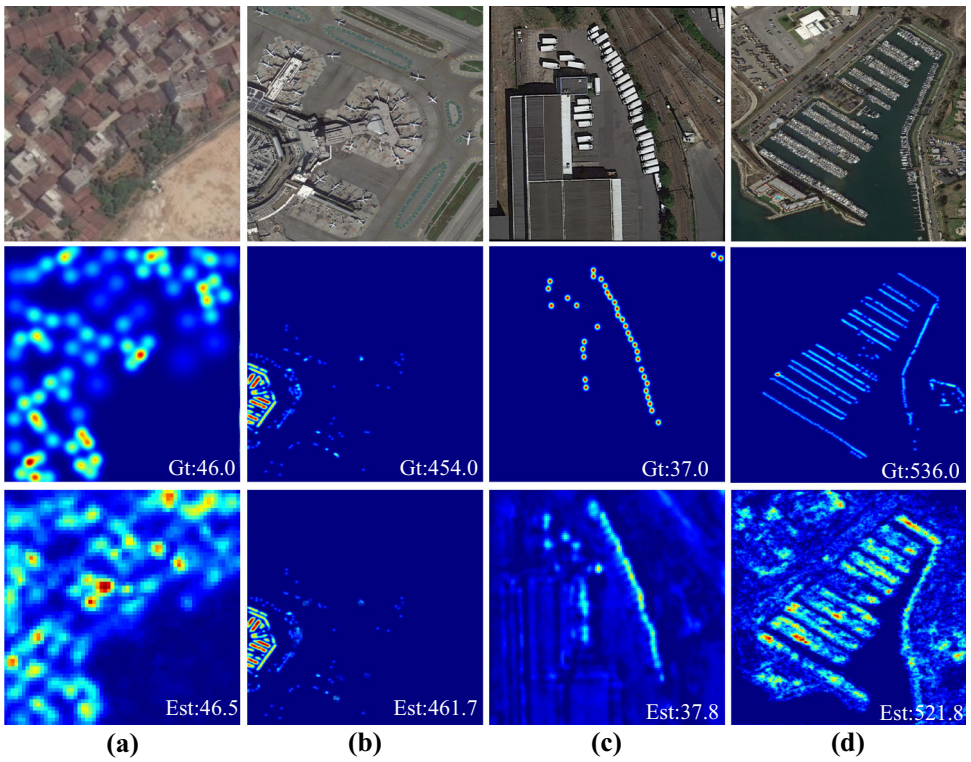

Fig. 6 Visualized results of the DAFNet on RSOC dataset. The top row denotes the input image, the middle row is the ground truth, and the bottom row shows the estimated density map. (a)∼(d) :The visualization of building, small-vehicle, large-vehicle and ship subdatasets, respectively. "Gt" and "Est" denote the ground truth and estimated count values, respectively

Table 2 Comparative results on the small-vehicle subdataset. The best performances are highlighted in **bold**

Methods	MAE	RMSE
baseline	498.85	1322.17
baseline+DSP	431.12	1278.35
baseline+IAF	424.42	1274.19
baseline+DSP+IAF	417.78	1240.54
baseline+IAF+DSP	**404.46**	**1211.78**

adding the DSP module to the baseline, the performance improves 13.6% and 3.3% in MAE and RMSE, respectively. Then, we add these two modules to the baseline in different orders. Adopting the DSP module first and the IAF module second on the basis of baseline, it scores 417.78 and, 1240.54 in MAE and RMSE, respectively. By comparison, it achieves the best counting performance, by adding the IAF module first and the DSP module second. Compared with the baseline, the MAE and RMSE decrease by 18.9% and 8.3%, respectively. As mentioned in Section 3.1, the IAF module is designed to suppress the background clutter and select a region where objects exist, while the DSP module is built to solve scale variation. In the configuration with the IAF module first and DSP second, the DSP module can be adopted to achieve more precise optimization in the region selected by the IAF module. However, when the order is reverse (i.e., DSP module first and IAF second), the result is suboptimal.

4.5 Failure cases

Despite the proposed DAFNet exhibits the superior experimental results against other mainstream methods, the counting performance is unsatisfactory on small-vehicle and ship subdatasets, as depicted in Table 1. In these two subdatasets, the small size of vehicles and ships makes it difficult to provide rich semantic information, which increases the difficulty of training models. In future work, we will study feature extraction algorithms for small objects to improve counting performance (Fig. 7).

5 Conclusion and future work

We propose a DAFNet for accurate object counting in remote sensing images. The IAF module is built to address the background clutter by emphasizing the region containing the objects. The key component of the IAF module is the MCA unit, which is built to merge the local and global features. To address the problem of scale variation in dense region, a DSP module is built, which adopts diverse dilated convolution layers with small dilated rates (1, 2 and 3) and improves the counting performance by capturing a large receptive field. Finally, we

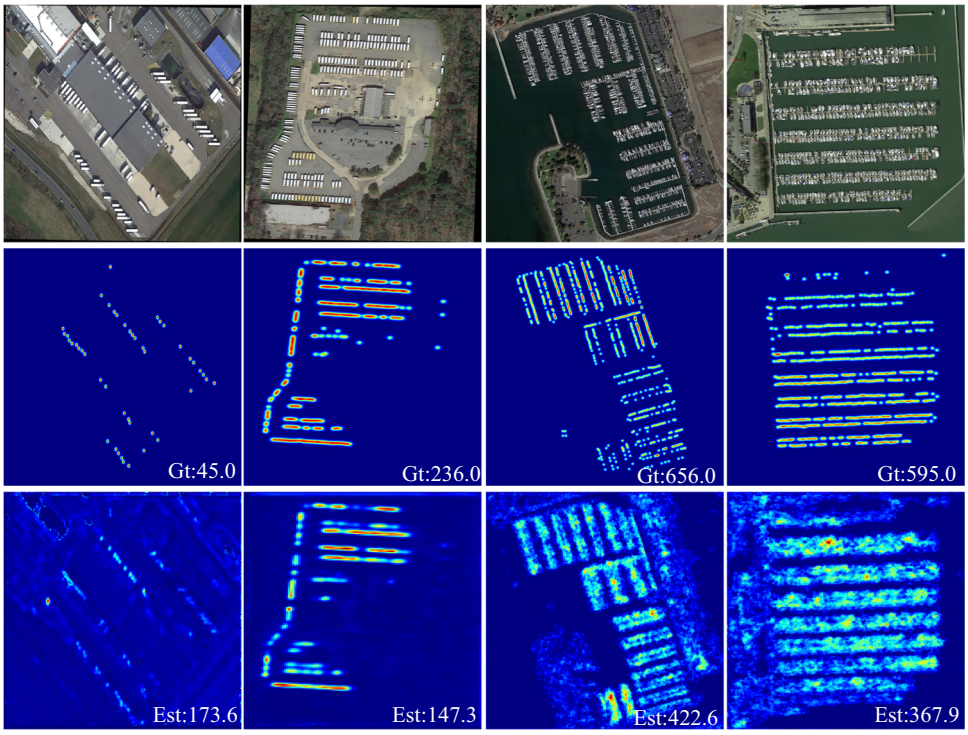

Fig. 7 Failure cases in RSOC dataset. The original, ground truth, and estimated images are represented from top to bottom. "Gt" and "Est" denote the ground truth and estimated count values, respectively

introduce the deformable convolution layers to handle the orientation arbitrariness. Experimental results prove that the DAFNet outperforms the state-of-the-art methods in remote sensing scenarios with background clusters and large-scale variations.

It is worth mentioning that the scores of MAE and RMSE on the small-vehicle and ship subdatasets are still unsatisfactory. In future work, we intend to design a better module to enhance performance against small objects counts.

Author Contributions **Xiangyu Guo**: Conceptualization, Methodology, Data Curation, and Writing - Original Draft. **Mingliang Gao**: Supervision, Formal analysis, Investigation, and Funding Acquisition. **Wenzhe Zhai**: Data Curation, Data Visualization, and Investigation. **Qilei Li**: Investigation, and Software. **Kyu Hyung Kim**: Formal Analysis, and Writing -Review & Editing. **Gwanggil Jeon**: Validation, and Writing -Review & Editing.

Funding This work is supported in part by the National Natural Science Foundation of China (Nos. 61601266 and 61801272) and National Natural Science Foundation of Shandong Province (Nos. ZR2021QD041 and ZR2020MF127).

Data Availability Not applicable.

Declarations

Ethics approval We declare that there is no ethics issue.

Conflict of Interests We declare that we have no conflict of interest.

References

1. Pallavi S, Mallapur JD, Bendigeri KY (2017) Remote sensing and controlling of greenhouse agriculture parameters based on iot. In: 2017 International conference on big data, IoT and data science (BID). IEEE, pp 44–48
2. Zhao W, Ma W, Jiao L, Chen P, Yang S, Hou B (2019) Multi-scale image block-level f-cnn for remote sensing images object detection. IEEE Access 7:43607–43621. https://doi.org/10.1109/ACCESS.2019.2908016
3. Cheng G, Si Y, Hong HDT, Yao X, Guo L (2021) Cross-scale feature fusion for object detection in optical remote sensing images. IEEE Geosci Remote Sens Lett 18:431–435. https://doi.org/10.1109/LGRS.2020.2975541
4. Kotaridis I, Lazaridou M (2021) Remote sensing image segmentation advances: a meta-analysis. Isprs Journal of Photogrammetry and Remote Sensing 173:309–322. https://doi.org/10.1016/J.ISPRSJPRS.2021.01.020
5. Xu Z, Zhang W, Zhang T, Yang Z, Li J (2021) Efficient transformer for remote sensing image segmentation. Remote Sens 13:3585. https://doi.org/10.3390/rs13183585
6. Rathore MM, Ahmad A, Paul A, Rho S (2016) Urban planning and building smart cities based on the internet of things using big data analytics. Comput Netw 101:63–80. https://doi.org/10.1016/j.comnet.2015.12.023
7. Pekel J-F, Cottam A, Gorelick N, Belward AS (2016) High-resolution mapping of global surface water and its long-term changes. Nature 540:418–422. https://doi.org/10.1038/nature20584
8. Fan Y, Wen Q, Wang W, Wang P, Li L, Zhang P (2017) Quantifying disaster physical damage using remote sensing data—a technical work flow and case study of the 2014 ludian earthquake in china. International Journal of Disaster Risk Science 8:471–488. https://doi.org/10.1007/s13753-017-0143-8
9. Gao J, Wang Q, Yuan Y (2019) Scar: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing 363:1–8. https://doi.org/10.1016/j.neucom.2019.08.018
10. Chen X, Bin Y, Sang N, Gao C (2019) Scale pyramid network for crowd counting. In: 2019 IEEE Winter conference on applications of computer vision (WACV), pp 1941–1950. https://doi.org/10.1109/WACV.2019.00211
11. Arteta C, Lempitsky V, Zisserman A (2016) Counting in the wild. In: European conference on computer vision. Springer, pp 483–498. https://doi.org/10.1007/978-3-319-46478-7_30
12. Loh DR, Yong WX, Yapeter J, Subburaj K, Chandramohanadas R (2021) 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 88:101845. https://doi.org/10.1016/j.compmedimag.2020.101845
13. Dai Z, Song H, Wang X, Fang Y, Yun X, Zhang Z, Li H (2019) Video-based vehicle counting framework. IEEE Access 7:64460–64470. https://doi.org/10.1109/ACCESS.2019.2914254
14. Topkaya IS, Erdogan H, Porikli FM (2014) Counting people by clustering person detector outputs. 313–318. https://doi.org/10.1109/AVSS.2014.6918687
15. Li M, Zhang Z, Huang K, Tan T (2008) Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1–4. https://doi.org/10.1109/ICPR.2008.4761705
16. Lempitsky VS, Zisserman A (2010) Learning to count objects in images. In: NIPS
17. Pham VQ, Kozakaya T, Yamaguchi O, Okada R (2015) Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In: Proceedings of the international conference on computer vision (ICCV), pp 3253–3261. https://doi.org/10.1109/ICCV.2015.372
18. Li M, Zhang Z, Huang K, Tan T (2008) Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1–4. https://doi.org/10.1109/ICPR.2008.4761705
19. Ge W, Collins RT (2009) Marked point processes for crowd counting. In: CVPR. https://doi.org/10.1109/CVPR.2009.5206621
20. Gao G, Gao J, Liu Q, Wang Q, Wang Y (2020) Cnn-based density estimation and crowd counting: A survey. arXiv:2003.12783
21. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 589–597. https://doi.org/10.1109/CVPR.2016.70
22. Li Y, Zhang X, Chen D (2018) Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1091–1100. https://doi.org/10.1109/CVPR.2018.00120
23. Liu W, Salzmann M, Fua P (2019) Context-aware crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5094–5103. https://doi.org/10.1109/CVPR.2019.00524

24. Gao J, Wang Q, Li X (2020) Pcc net: Perspective crowd counting via spatial convolutional network. IEEE Trans Circuits Syst Video Technol 30:3486–3498. https://doi.org/10.1109/TCSVT.2019.2919139
25. de Santana Correia A, Colombini E (2021) Attention, please! a survey of neural attention models in deep learning. arXiv:2103.16775
26. Zhai W, Li Q, Zhou Y, Li X, Pan J, Zou G, Gao M (2022) Da2net: A dual attention-aware network for robust crowd counting Multimedia Systems PP. https://doi.org/10.1007/s00530-021-00877-4
27. Zhai W, Gao M, Anisetti M, Li Q, Jeon S, Pan J (2022) Group-split attention network for crowd counting. Journal of Electronic Imaging. https://doi.org/10.1117/1.JEI.31.4.041214
28. Liu N, Long Y, Zou C, Niu Q, Pan L, Wu H (2019) Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3225–3234. https://doi.org/10.1109/CVPR.2019.00334
29. Sindagi VA, Patel VM (2020) Ha-ccn: Hierarchical attention-based crowd counting network. IEEE Trans Image Process 29:323–335. https://doi.org/10.1109/TIP.2019.2928634
30. Jiang X, Zhang L, Xu M, Zhang T, Lv P, Zhou B, Yang X, Pang Y (2020) Attention scaling for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4705–4714. https://doi.org/10.1109/cvpr42600.2020.00476
31. Rong L, Li C (2021) Coarse- and fine-grained attention network with background-aware loss for crowd density map estimation. In: Proceedings of the IEEE workshop on applications of computer vision (WACV), pp 3674–3683
32. Gao G, Liu Q, Wang Y (2021) Counting from sky: A large-scale data set for remote sensing object counting and a benchmark method. IEEE Trans Geosci Remote Sens 59:3642–3655. https://doi.org/10.1109/TGRS.2020.3020555
33. Sindagi V, Patel V (2017) Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: Proceedings of the IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–6. https://doi.org/10.1109/AVSS.2017.8078491
34. Cao X, Wang Z, Zhao Y, Su F (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European conference on computer vision (ECCV). https://doi.org/10.1007/978-3-030-01228-1_45
35. Wang Q, Gao J, Lin W, Yuan Y (2019) Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 8190–8199. https://doi.org/10.1109/CVPR.2019.00839
36. Zhu L, Zhao Z, Lu C, Lin Y, Peng Y, Yao T (2019) Dual path multi-scale fusion networks with attention for crowd counting. arXiv:1902.01115