Pattern Recognition 126 (2022) 108548

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions[☆]

Vishnu Chalavadi[a,*], Prudviraj Jeripothula[a], Rajeshreddy Datla[a,b], Sobhan Babu Ch[a], Krishna Mohan C[a]

[a] Visual Learning and Intelligence Group (VIGIL), Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Kandi, Sangareddy 502285, India
[b] Advanced Data Processing Research Institute (ADRIN), Department of Space, Akbar Road, Tarbund, Manovikas Nagar, Secunderabad-500009, India

ARTICLE INFO

Article history:
Received 4 October 2021
Revised 22 December 2021
Accepted 21 January 2022
Available online 23 January 2022

Keywords:
Multi-scale object detection
Contextual features
Dilated convolutions
Aerial images

ABSTRACT

The object detection in aerial images is one of the most commonly used tasks in the wide-range of computer vision applications. However, the object detection is more challenging due to the following issues: (a) the pixel occupancy vary among the different scales of objects, (b) the distribution of objects is not uniform in aerial images, (c) the appearance of an object varies with different view-points and illumination conditions, and (d) the number of objects, even though they belong to same type, vary across the images. To address these issues, we propose a novel network for multi-scale object detection in aerial images using hierarchical dilated convolutions, called as mSODANet. In particular, we probe hierarchical dilated network using parallel dilated convolutions to learn the contextual information of different types of objects at multiple scales and multiple field-of-views. The introduced hierarchical dilated network captures the visual information of aerial image more effectively and enhances the detection capability of the model. Further, the extensive experiments conducted on three challenging publicly available datasets, i.e., Visdrone2019, DOTA (OBB & HBB), NWPU VHR-10, demonstrate the effectiveness of the proposed mSODANet and achieve the state-of-the-art performance on all three datasets.

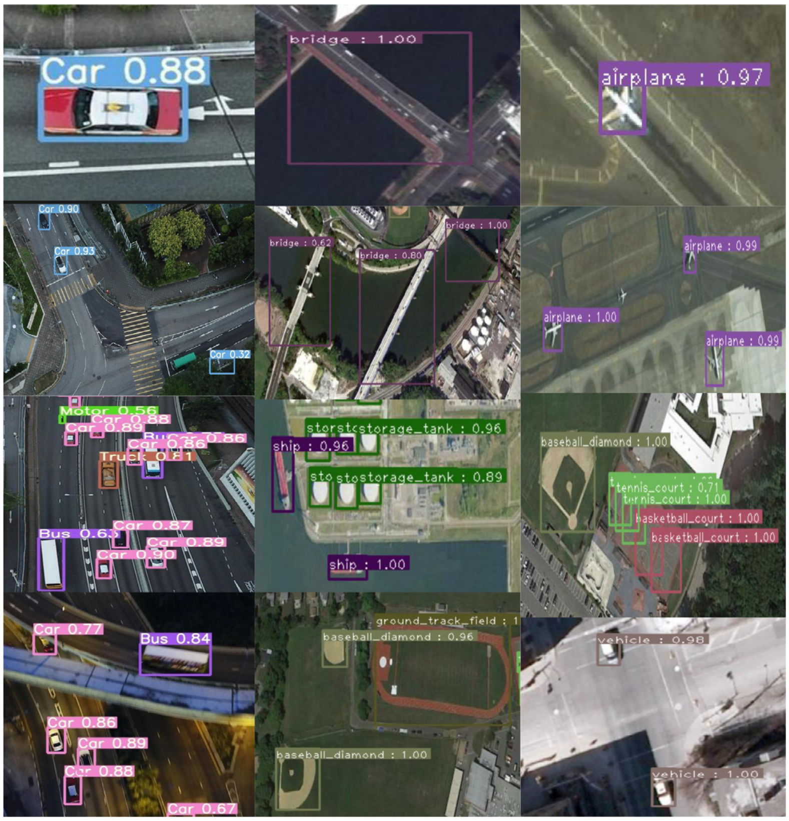
1. Introduction

The content in aerial images encompasses various characteristics of objects on the ground due to non-uniformity in the acquisition. These characteristics are manifested differently and specific to the acquisition platforms like drones and airplanes. Especially with aerial images, many real-world vision based applications require the knowledge of their characteristics, e.g., depiction of objects at different scales, 0 to 360 degree top-view, non-uniformity in their spatial arrangements, and different contextual background information, etc., for better analysis. Particularly, object detection is the process of individual object localization and its classification that requires a thorough understanding of the characteristics of aerial images. Moreover, the objects irrespective of their types, have different backgrounds and vary seasonally as well with view-angle of the sensor. Thus, the background information of objects indicates the contextual information, which provides useful semantics to the detection process. Hence, the identification of an object type at various scales along with the presence of other objects and the background information is quite difficult and challenging. For example, *bridge* object in aerial image can have different background information, i.e., either sand or water or both. In addition to different backgrounds of the same object, the pixel occupancy of the background is more than the foreground resulting in high intra-class variation. Also, the pixel occupancy of larger objects (e.g., *ground-track field*) is more than smaller objects (e.g., *ship, airplane*) in aerial images. Some objects such as *basketball court* and *tennis court* look visually similar and their uniform background information causes low inter-class variation. Fig. 1 depicts different objects and their characteristics in aerial images, besides their intra-class and inter-class variations. Hence, the object detection model must comprehend these characteristics, which jointly constitute the inherent properties of different objects in aerial images.

Existing convolutional neural network (CNN) based approaches [1,2] have proven their effectiveness in improving the performance of object detection task, especially in natural images (e.g., Pascal VOC [3] and MS COCO dataset [4]). These approaches are further improved by employing two CNN models separately to learn rotation-invariant features, besides imposing Fisher discrimination criterion in the objective functions of CNN model to extract effec-

[☆] Source code is available at https://github.com/cv15nu/mSODANet
[*] Corresponding author.
E-mail address: cs16m18p000001@iith.ac.in (V. Chalavadi).

https://doi.org/10.1016/j.patcog.2022.108548
0031-3203/© 2022 Elsevier Ltd. All rights reserved.



(a) VisDrone2019 (b) DOTA (c) NWPU VHR-10

Fig. 1. Characteristics of different objects from three varieties of aerial imagery object detection datasets. Row 1 & Row 2 depict the similar objects at multiple scales causing large disparity in their pixel occupancy in the images. Row 3 shows the densely packed and arbitrary sized objects. Row 4 includes the objects with arbitrary spatial arrangements.

tive CNN features [5]. However, these approaches are not adequate in understanding the special characteristics of aerial images in order to achieve better performance in object detection task. Unlike natural images, aerial images have different characteristics such as (i) view-point variations while acquisition, (ii) similar objects with multiple scales, (iii) the large deviation in pixel occupancy from small, medium to large objects, (iv) complex background information of the objects, and (v) non-uniform distribution of objects along with arbitrary spatial arrangements. To analyze such characteristics, many works [6,7] have been proposed on multi-scale object detection in aerial images. For instance, Zhang et al. [6] introduced a context aware detection network (CAD-Net) which primarily uses ResNet [8] & feature pyramid network (FPN) [9] as backbone feature networks and constructs global & local contexts of objects. However, CAD-Net is unable to learn contextual information of objects effectively, at multiple scales and multiple filed-of-views. Recently, Li et al. [7] proposed ground sample distance identifying network (GSDet) based on dilated convolutions to utilize ground sample distance information and further enhance detection capability. Even though the GSDet network leverages dilated convolutions to process objects at multiple scales & incorporates larger receptive field, it lacks the feature refine module which effectively distills the multi-scale features.

To address the aforementioned challenges, in this work, we propose a novel multi-scale object detection network, mSODANet, using hierarchical dilated convolutions to learn contextual features and model effective object detection framework. In particular, we first extract multi-scale features from EfficientNet backbone network and then construct the hierarchical dilated network (HDN) in order to extract contextual features at multiple-scales and multiple-field of views. In addition, we explore bi-directional feature aggregation module (BFAM) to further refine the obtained multi-scale features and achieve dense multi-scale contextual features. Thus, we enhance the detection capability of the proposed model and achieve significant improvement over recent state-of-the-art models [6,7,10]. The main contributions of this paper are summarized as

- We present a novel framework, mSODANet for effective multi-scale object detection in aerial images.
- To cope with arbitrary size objects in aerial images, we learn multi-scale contextual information using hierarchical dilated convolutions.
- Extensive experiments are conducted on three challenging datasets, i.e., VisDrone2019, DOTA (OBB & HBB), and NWPU VHR-10 to validate the efficacy of mSODANet.

2. Related work

This section provides a brief review of the existing methods on object detection in aerial images. Most of the CNN based object detection methods have been evolved using region based convolutional neural networks [2], which are designed for natural images. In [11,12], different types of objects are detected by accommodating the multiple scales along with 0–360 degrees rotation of the objects from top-view. These multiple rotations of geo-spatial objects in aerial images are accommodated by incorporating a layer on the existing CNN architectures [11]. A feature pyramid network (FPN) [9] is developed as a top-down architecture by employing lateral connections to build feature maps of high-level semantics at different scales. The contextual background information of objects in aerial images vary at multiple scales. A multi-scale detection framework [13] is used to generate quality proposals in order to detect multi-scale objects effectively.

Some methods [8,14] focus on capturing the arbitrary sizes of different objects in aerial images. SNIPER [14] exhibits promising results among these methods. Also, frameworks with patch-based training have been considered for efficient object detection [15]. A one-stage network (AVDNet) [16] is designed to detect the small-sized objects. In AVDNet, residual blocks at multiple scales are introduced to preserve the vanishing features for smaller objects. The residual blocks along with output feature map achieves an effective representation of the salient features of the small objects. Also, a visualization mechanism for recurrent-features (RFAV) is used to analyze the network behavior.

The localization scheme is very crucial along with classification accuracy in order to detect the densely packed objects with different orientations. To detect the objects effectively in densely packed arrangements, an RoI transformer [17] is used to transform the region of interest from horizontal to rotational. This transformer tackles the misalignment problem between the localization accuracy and classification confidence. Similarly, an end-to-end framework [18] is designed by unifying the objects as clusters for their detection in aerial images. This framework comprises of three subnetworks: cluster proposal, scale estimation, and detection tasks. The cluster proposals for a given input image are generated and these proposals are used to estimate the object scales. Then, the normalized clusters are fed into the detection network. This framework effectively minimizes the number of patches in the detection process and also helps to boost detection performance. A SCRNet [19] is designed to detect the small-sized and rotated objects in the cluttered scenes. Here, multi-layer features are fused effectively using anchors to increase the sensitivity to small-sized objects. Both the channel attention and supervised pixel attention networks help to mitigate the presence of noise and retain the relevant features of small and cluttered objects.

In [20], “random access memories” approach is proposed for target detection, instead of localization and classification schemes at inference stage. This approach is formulated from a Bayesian perspective, in which the detection model is updated adaptively to maximize its posterior using both training and observation samples. To improve the performance of existing state-of-the-art object detection methods, an approach [5] imposes both rotation-invariant and a Fisher discrimination regularizers on CNN features. A density-map guided network [21] is designed to obtain the distribution of objects in feature map with respect to the change in pixel intensity. The intensity variations sense the existence of objects and thereby provide statistical guidance to crop the images. This network has three modules, i.e., density map generation, image cropping, and detector. The generated density maps are helpful to learn the scale information in order to crop the regions in images. The class imbalance in the detection process is addressed using network [22] with dual sampler (DSHNet) to resolve the issues with long-tail distribution in aerial images. Recently, ground sample distance (GSD) of aerial images is incorporated in modeling the object detection process [7]. This is a two-stage detection framework, which has a subnet to convert the GSD regression into a probability estimates. Subsequently, the GSD information is combined with the sizes of regions of interest to determine the actual size of objects. However, the efficacy of this two-stage detection framework [7] has been investigated and its performance is reported only on DOTA (Oriented Bounding Box) dataset.

Recently, context-aware approaches is used to achieve multi-scale object detection of small objects. Spatio-temporal convnet (STDnet-ST) [23] exploits the correlation of promising regions between frames for those objects under 16×16 pixels using an efficient tubelet linking is performed to link small objects across video frames. Multi- scale Structural Kernel Representation (MSKR) [24] employs polynomial kernel approximation, which does not only draw into high-order statistics but also preserve the spatial information of input. Multi-scale deep feature learning network (MDFN) [25] efficiently detects the objects by introducing information square and cubic inception modules into the high-level layers, which employs parameter-sharing to enhance the computational efficiency. It considers not only individual objects and local contexts but also their relationships. Spatial Context-aware Network (SCA-Net) [26] adopts a Short-Path Context Module (SPCM) to progressively enforce the interaction between local contextual cues and global features.

Further, the attention modules are used to enhance the multi-scale object detection performance in a deeply supervised U-like encoder-decoder network [27], which consists of feature extraction, channel-wise attention, boundary information localization and saliency fusion modules. In order to obtain multi-scale contextual information, Content-Aware Guidance Network (CAGNet) [28] utilizes a Multi-scale Feature Extraction Module (MFEM) at each level of abstraction. Finally, a hybrid loss function is designed w.r.t scale-balanced loss [29] which outperforms the widely used Cross-entropy loss. Specifically for remote sensing images, Contrast-weighted Dictionary Learning (CDL) [30] is proposed to learn salient and non-salient atoms from positive and negative samples to construct a discriminant dictionary, in which a contrast-weighted term is proposed to encourage the contrast-weighted patterns to be present in the learned salient dictionary while discouraging them from being present in the non-salient dictionary.

Different from the existing approaches, we attempt to exploit effective multi-scale contextual information for multi-scale object detection by jointly combining the three different networks, i.e., backbone network, hierarchical dilated network, bi-directional feature pyramid network. Moreover, the existing methods followed various approaches that consider the characteristics of the objects of either VisDrone2019 or DOTA or NWPU VHR-10 datasets or any two of them. In this paper, we focus on a joint network for object detection to investigate all three varieties of aerial imagery object detection datasets.

3. Proposed method

In this section, we present mSODANet, a novel network for multi-scale object detection in aerial images. The complete framework of the proposed model is illustrated in Fig. 2. The proposed mSODANet has three major components, backbone network, hierarchical dilated network (HDN), and bi-directional feature aggregation module (BFAM). The backbone network (Section 3.1) presents the feature extraction process. The HDN (Section 3.2) learns the multi-scale contextual information. And, the BFAM (Section 3.3) refines the attained multi-scale contextual features for effective multi-scale object detection.

3.1. Backbone network

Aiming at robustness and efficiency, we use recent state-of-the-art image classification model, EfficientNet [31], as a backbone network for our multi-scale object detection framework where it scales up the dimensions of network depth, width, and input resolution. Mainly, we leverage the Imagenet pretrained checkpoints from different scaling coefficients of EfficientNet-B0 to B6 as a backbone feature extractor in order to fully utilize the visual information of varied size input images. In addition, we incorporate compound scaling to jointly scale up the whole object detection framework as in Tan et al. [32] to achieve better performance. On extracting image features from EffecientNet, we forward them to

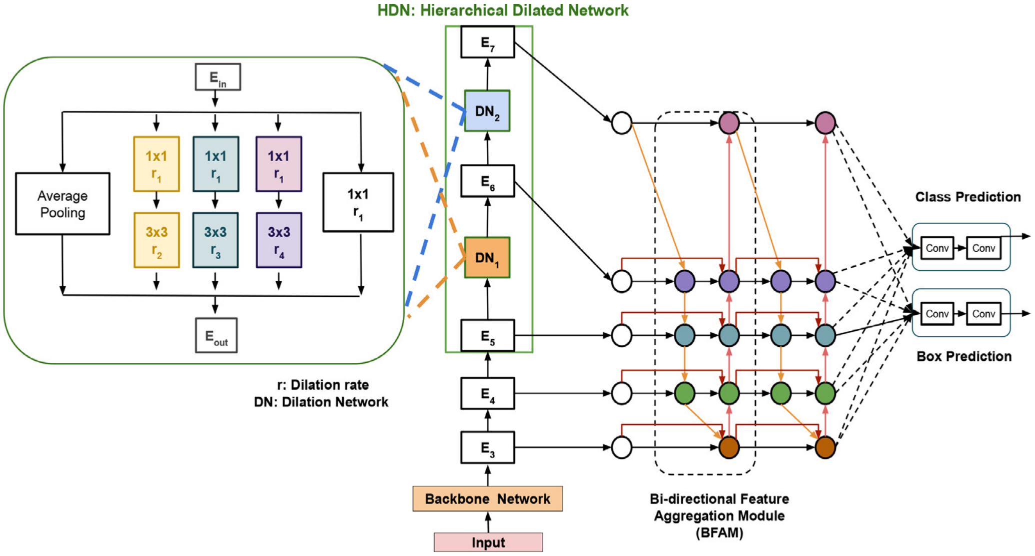


Fig. 2. Block diagram of the proposed hierarchical dilated network for object detection in aerial images.

hierarchical dilated network (HDN) to learn the multi-scale contextual features of arbitrary size objects.

3.2. Hierarchical dilated network (HDN)

Usually, aerial images cover large visual scene that encloses rich contextual information of multiple objects. The contextual information of visual objects characterizes the correlation between semantic nature of visual object and visual scene. In addition, the scale of objects in visual scene varies drastically over the all aerial images. Hence, it is essential to develop a model that can handle varied size objects and capture the contextual information of all the objects present in a visual scene. Motivated by the above observations, we learn the contextual information of various objects at multiple-scales and multiple field-of-views by introducing hierarchical dilated network on top of multi-scale features extracted from backbone network. The proposed hierarchical dilated network component is shown in Fig. 2 (left), where we incorporate multiple parallel atrous/dilated convolutions in each level of HDN.

On extracting multi-scale features from backbone network [$\vec{E}^{in} = E^{in}_{l_1}, E^{in}_{l_2}, \ldots$], where $E^{in}_{l_i}$ denotes the feature at level l_i, we define a transformation function h such that the all extracted features are aggregated effectively to produce the output features $\vec{E}^{out} = h(\vec{E}^{in})$. Commonly, by taking input features at levels from 3 to 7, i.e., $\vec{E}^{in} = E^{in}_3, \ldots, E^{in}_7$, where E represents the feature map of size $1/2^i$, the recent object detectors like FPN [9] or EfficientDet [32] aggregate multi-scale features and produce output features as

$$E^{out}_7 = Conv(E^{in}_7)$$

$$E^{out}_6 = Conv(E^{in}_6 + G(E^{out}_7))$$

$$\ldots$$

$$E^{out}_3 = Conv(E^{in}_3 + G(E^{out}_4)), \tag{1}$$

where G indicates the upsampling or downsampling operation and $Conv$ represents the typical convolutional operation. Even though, the above feature fusion mechanism incorporates the multi-scale information, it fails to learn multi-scale contextual information of all objects in an visual scene. Also, standard convolutions, due to their small receptive field, may not capture variations in the object scales and perspective effectively.

To address aforementioned issues, in our work, we introduce the hierarchical dilated network (HDN) to effectively capture contextual information at multiple scales and multiple filed-of-views. Our hierarchical dilated network introduces two dilated networks (DN_1 and DN_2) in between multi-scale features which are extracted from backbone network. As shown in Fig. 2, we first extract multi-scale features (E_3, E_4, E_5, E_6, E_7) using pretrained EfficientNet network and then employ two dilated networks DN_1 and DN_2 in between E_5 & E_6 and E_6 & E_7 in order to construct hierarchic dilated network. Both DN_1 and DN_2 of HDN incorporate parallel dilated convolutions with different dilation rates to capture larger receptive field information and further incorporate multi-scale contextual information. Dilated convolutions are extremely powerful convolutional operations which explicitly control the receptive field of feature maps and process the input at the multiple field-of-views [33].

Given a two dimensional signal, for each location p on the output convolutional feature map g and filter w, the dilated convolution is applied on input f as

$$g[p] = \sum_j f[p + r \cdot j]w[j], \tag{2}$$

where r is the dilation rate which determines the stride that we sample from the input signal. Note that the $r = 1$ is the standard convolution, a special case of dilated convolutions.

In both DN_1 and DN_2 of our hierarchical dilated network, we employ several parallel dilated convolutions with different dilation rates and filter sizes. The construction of DN_1 and DN_2 is shown in Fig. 2 (left). Here, both DN_1 & DN_2 follow the same architectural design, where we incorporate several $f \times f$ filters with dilation rates of r_1, r_2, r_3, r_4. On incorporating DN_1 in between E_6 & E_7 and DN_2 in E_5 & E_6, the Eq. (1) is updated to

$$E^{out}_7 = Conv(E^{out}_6)$$

$$E^{out}_6 = DN_2(E^{out}_5 + R(E^{out}_7))$$

$$E^{out}_5 = DN_1(E^{in}_5 + R(E^{out}_6))$$

$$\ldots$$

$$E^{out}_3 = Conv(E^{in}_3 + R(E^{out}_4)), \tag{3}$$

where DN_1 & DN_2 are the dilated network components of HDN.

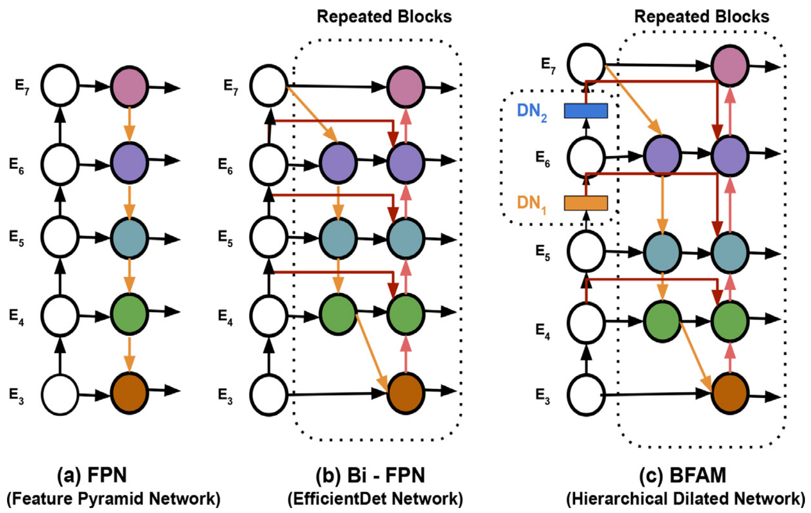


Fig. 3. Feature network comparison.

3.3. Bi-directional feature aggregation module (BFAM)

The detection framework with feature pyramid network (FPN) [9] has been extensively investigated for multi-scale object detection task [6,34]. However, the FPN [9] architecture is inherently restricted to top-down information only. To avoid such issue, Liu et al. introduced a path aggregation network (PANet) [35] by adding an extra bottom-up pathway to the existing top-down pathway as shown in Fig. 3(b). Further, NAS-FPN [36] explored neural architecture search (NAS) in order to search for efficient cross-scale feature network topology. Recently, a bi-directional feature pyramid network (Bi-FPN) is introduced in Tan et al. [32] with several optimizations on cross-scale connections in order to improve the model efficacy and further achieve state-of-the-art performance. Inspired by this, we refine the extracted multi-scale contextual features using BFAM as shown in Fig. 3(c), where we first extract the multi-scale contextual features at various levels from 3 to 7 $\{E_3, E_4, E_5, E_6, E_7\}$ as formulated in Eq. (3) and recurrently apply the feature aggregation mechanism.

In brief, the bi-directional feature aggregation module (BFAM) operates at three stages. At first, the BFAM discards the nodes with which it has only one input edge on the basis that the single edge nodes does not contribute much information into the network. Second, it adds an extra edge from the original input feature map to each output feature map at the same level to combine more information without adding any learnable layers. Third, it integrates top-down and bottom-up path ways as one feature layer of BFAM, and further employs such layers recurrently to learn dense multi-scale contextual information. In our work, we process the multi-scale contextual information extracted from hierarchical dilated network by the BFAM layers at multiple times to establish effective multi-scale contextual information within the network and further enhance the object detection capability. Moreover, we follow the same box prediction process as in BFAM in order to generate the predictions of class and bounding box regressor, since the BFAM is employed on top of hierarchical dilated network module.

4. Experimental results

This section presents both quantitative and qualitative analysis of the proposed method to demonstrate its effectiveness on three varieties of benchmark object detection datasets in aerial images, namely, VisDrone2019 [37], DOTA (OBB & HBB) [38], and NWPU VHR10 [39].

4.1. Datasets

VisDrone2019 It is the largest drone dataset which consists of 8599 images with 10-class objects comprising 6471 images for training, 548 for validation, and 1580 for testing. The objects from 10 classes are *Awning, Bicycle, Bus, Car, Motor, Pedestrian, People, Truck, Tricycle*, and *Van*.

DOTA It is the largest aerial image object detection dataset which is available in two variants based on Horizontal and Oriented bounding-box annotations, i.e., DOTA (HBB) & DOTA (OBB). This DOTA dataset has 2806 total images with 188, 282 instances depicting 15 categories of objects. These categories include *Baseball diamond (BD), Basketball court (BC), Bridge (BR), Ground track field (GTF), Harbor (HA), Helicopter (HC), Large vehicle (LV), Plane (PL), Roundabout (RA), Ship (SH), Small vehicle (SV), Storage tank (ST), Soccer-ball field (SBF), Swimming pool (SP)*, and *Tennis court (TC)*. The instances have large variations in scale, orientation, and aspect ratio. Both DOTA (HBB) & DOTA (OBB) datasets are split into 1411 images for training, 458 image for validation, 937 images for testing. *NWPU VHR-10* It consists of 650 remote sensing images comprising 10 categories of objects. These categories include *Airplane, Storage tank, Ship, Baseball diamond, Basketball court, Tennis court, Harbor, Ground track field, Bridge*, and *Vehicle*. The split ratio of this dataset for train, validation, and test is 80%,10%, and 10% ratios, respectively.

Each of the datasets consists of multiple objects acquired from platforms with different sensing parameters such as view-point, height of the acquisition platform, etc., which influence the depiction of various objects of same type. Thus the objects in VisDrone2019 have a wide range of scale variations, illumination conditions, scenarios, and view-points. Similarly, the objects in DOTA are densely packed and also exhibit high inter-class and low intraclass variations. The objects in NWPU VHR-10 dataset are at different scales with complex background information. The pixel occupancy of the objects in these images are influenced by the scale of the object(s) and the number of objects in an image. Hence, these datasets together pose various challenges in designing a single stage object detection framework to incorporate their characteristics jointly in the detection process.

4.2. Implementation details

Data augmentation To effectively utilize the characteristics of aerial images to detect objects, multiple data augmentation operations are explored to improve the training in the proposed

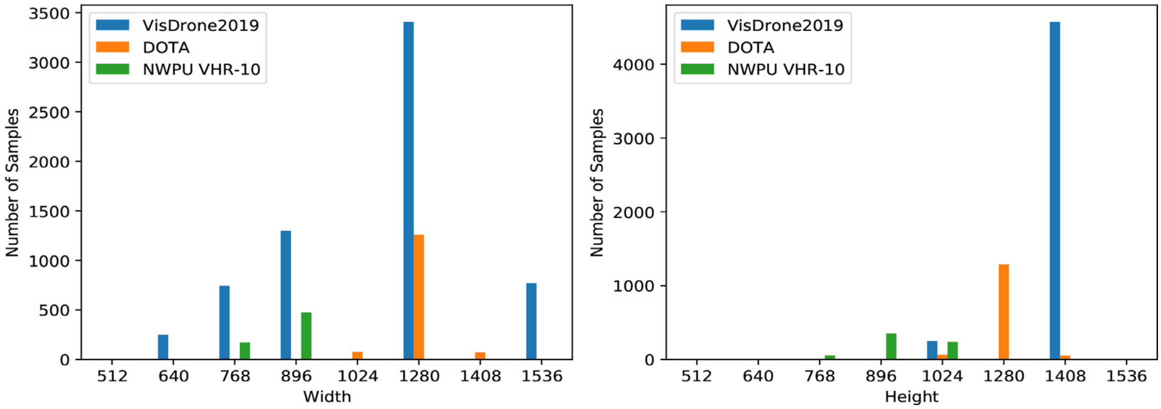


Fig. 4. Dimensions of input image samples from three object detection datasets.

method. Specifically, vertical flipping, horizontal flipping, rotation from -90° to 90°, scaling from 0.5 to 1.5, and shear range -16 to 16. The adoption of data augmentation in our pipeline contributes to the robustness of the proposed model under varying geometric features.

Proposed multi-scale object detection model For our mSODANet model, we adopt the open source implementation of EfficientNet [31]. And, we reuse the pre-trained checkpoints of EfficientNet (B_3 and B_5) as backbone feature extractor. Usually, the EfficientNet provides multiple checkpoints from B_0 to B_7 for effective representation of various image resolutions. In this work, we choose the suitable checkpoints by generating a 2D histogram consisting of width and height for all the three benchmark object detection datasets. The dimension (width & height) in majority of the images of VisDrone2019, DOTA, and NWPU VHR-10 are between 1100–1400, 1200–1300, 750–900, respectively, as depicted in Fig. 4. Based upon this, we fix the EfficientNet backbone checkpoint as B_5 (1280) for VisDrone2019, DOTA OBB, & DOTA HBB, and B_3 (896) for NWPU VHR-10 dataset. In the hierarchical dilation network, we introduce dual dilation modules DN_1 & DN_2 consisting of dilation rates [1,2,4,6] & [1,6,8,10], respectively. The dilated modules use 1×1 and 3×3 dilated convolutions and 64 channels for parallel dilated convolutions. In addition, the object detection model is trained with a learning rate of 0.0003 and weight decay of $4\exp^{-5}$ using an SGD optimizer. Synchronized batch norm is utilized to calculate focal loss after every convolutional layer.

4.3. Quantitative analysis

This section presents the performance comparison of our proposed method with state-of-the art approaches on three publicly available aerial image object detection datasets, namely, VisDrone2019, DOTA (OBB & HBB), and NWPU VHR-10. We consider mean average precision (*mAP*) metric, specifically for the threshold of 0.5 & 0.75, to indicate a predicted bounding box if its Intersection over Union (IoU) is greater than 0.5 or 0.75. Also, we provide the average precision (*AP*) over small, medium, and large bounding box scales. We report the object detection performance of the proposed method on three aerial image datasets in the following sub-sections.

4.3.1. VisDrone2019

Table 1 gives threshold-wise object detection performance and Table 2 presents class-wise object detection performance of the proposed method. Our proposed method achieves a margin of 7.6% improvement in the overall threshold-wise average precision over the existing state-of-the art approaches. The threshold-wise metrics AP_{75}, AR_{10}, AR_{100}, and AR_{500} indicate the superiority of our method. Also, the class-wise average precision of our method outperforms state-of-the-art methods. These results signify the robustness of our mSODANet and its ability to encode the characteristics of even more harder samples in the detection process. This ensures that our mSODANet with hierarchical dilated convolutions effectively captures multi-scale objects with complex background information in variety of scenarios in aerial images.

4.3.2. DOTA

DOTA dataset provides two kinds of bounding boxes, i.e., DOTA oriented bounding boxes (OBB) and DOTA horizontal bounding boxes (HBB) for object detection task. In this work, we evaluate the performance of our mSODANet on both DOTA (OBB) and DOTA (HBB) datasets in order to demonstrate the robustness of the object detection capability. Tables 3 and 4 present the performance comparison of mSODANet with state-of-the-art methods in terms of average precision (AP) and threshold-wise average precisions (AP_{50} & AP_{75}) on DOTA (OBB) & DOTA (HBB) datasets, respectively. It is evident from the Tables that our mSODANet shows an improvement of 2% on DOTA (OBB) and 5% on DOTA (HBB) over the existing state-of-the-art methods in terms of average precision (AP). Thus we report the performance of our mSODANet as new state-of-the-art. In addition, the class-wise performance of the proposed mSODANet has been evaluated over state-of-the-art methods on both DOTA (OBB) and DOTA (HBB) in terms of mean average precision (*mAP*) as given in Tables 5 and 6, respectively. It can be observed from Table 5 that the proposed mSODANet outperforms on most of the classes in DOTA (OBB) dataset. For DOTA (HBB), we significantly outperform all the classes in terms of mean average precision (*mAP*) as shown in Table 6.

4.3.3. NWPU VHR-10

Table 7 gives the overall class-wise average precision of our proposed method and state-of-the art object detection approaches on NWPU VHR-10 dataset. The efficacy of mSODANet is compared with five different object detection frameworks. A margin of 4.5% improvement in the average precision of mSODANet is observed in comparison with state-of-the-art methods as shown in Table 7. Moreover, the detection performance of our mSODANet exceeds 97% on all objects except ship. Thus our method exhibits a consistent detection performance over all the individual objects in line with state-of-the art approaches. Specifically, our method exhibits an average of 8% significant improvement in *mAP* for objects such as *Tennis court, basketball court, bridge*, and *vehicle*. This shows that our method is able to encode the characteristics of different types of objects at multiple scales in addition to their contextual information effectively. It can be observed from Table 7 that the proposed method significantly exceeds the existing state-of-the-art

Table 1
Threshold-wise average precision (AP) comparison of mSODANet with state-of-the-art on VisDrone2019 dataset.

Method	AP (%)	AP_{50} (%)	AP_{75} (%)	AR_1 (%)	AR_{10} (%)	AR_{100} (%)	AR_{500} (%)
SAMFR-Cascade RCNN [40]	20.18	40.03	18.42	0.46	3.49	21.6	30.82
EfficientDet (B5) [32]	21.40	38.60	20.20	0.59	04.12	22.38	31.75
DSHNet [22]	24.60	44.40	24.10	–	–	–	–
RR-Net [41]	29.13	55.82	27.23	1.02	8.50	35.19	46.05
Patch Level Augmentation Net [42]	29.13	54.07	27.38	0.32	1.48	9.46	44.53
mSODANet - E_6, E_7 (Ours)	**36.89**	**55.92**	**37.41**	**1.15**	**11.36**	**37.25**	**48.92**

Table 2
Class-wise average precision (AP) comparison of mSODANet with state-of-the-art on VisDrone2019 dataset.

Method	mAP (%)	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning	Bus	Motor
EfficientDet (B5) [32]	21.40	19.30	13.10	9.60	44.80	29.40	21.30	13.20	8.50	35.60	19.40
DSHNet [22]	24.60	22.50	16.50	10.10	52.80	32.60	22.10	17.50	8.80	39.50	23.70
RR-Net [41]	29.13	30.44	14.85	13.72	51.42	36.14	35.22	28.01	18.99	44.20	25.85
mSODANet - E_6, E_7 (Ours)	**36.89**	**38.72**	**21.43**	**20.28**	**59.61**	**44.32**	**41.73**	**35.75**	**24.67**	**49.93**	**32.52**

Table 3
Threshold-wise average precision (AP) comparison of mSODANet with state-of-the-art methods on DOTA (OBB) dataset.

Method	AP (%)	AP_{50} (%)	AP_{75} (%)
RoI Transformer [17]	69.56	–	–
GSDet [7]	68.28	–	–
EfficientDet (B5) [32]	72.52	75.86	73.36
SCRDet [19]	72.61	–	–
HSP-Net [10]	80.42	–	–
mSODANet - E_6, E_7 (Ours)	**82.66**	**86.48**	**81.13**

Table 4
Threshold-wise average precision (AP) comparison of mSODANet with state-of-the-art methods on DOTA HBB dataset.

Method	AP (%)	AP_{50} (%)	AP_{75} (%)
CAD-Net [6]	69.9	–	–
SAPNet [43]	62.9	-	-
EfficientDet (B5) [32]	74.70	80.30	78.60
SCRDet [19]	75.35	–	–
AVDNet [16]	79.65	–	–
HSP-Net [10]	80.42	–	–
mSODANet - E_6, E_7 (Ours)	**85.83**	**90.33**	**89.61**

performance in most of the classes. This indicates the effectiveness of hierarchical dilated convolutions in mSODANet to capture the complex contextual information of different types of objects. However, the pixel occupancy of the contextual information accompanied with the small objects is relatively less than the object itself, which shows slightly less impact in *ship* class.

From the quantitative results, we observe that the mSODANet achieves a significant improvement in the detection performance over the recent methods, such as GSDet [7], CADNet [6], and HSP-Net [10], though they leverage the dilated convolutions, feature aggregation module, and their combination. This is due to the benefit of hierarchical dilated convolutions which are introduced at various levels of BFAM along with EfficientNet backbone network. Our mSODANet exhibits its consistency in handling the characteristics of all the objects in three object detection datasets.

4.4. Qualitative analysis

Fig. 5 illustrates the results of object detection on three varieties of benchmark datasets using our mSODANet. Fig. 5(a) presents the results of mSODANet on VisDrone2019 dataset depicting different types of objects under various scenarios. Similarly, the objects of NWPU VHR-10 dataset with different characteristics are detected using mSODANet and their results are shown in Fig. 5(b). And the results of mSODANet on different types of objects from both DOTA (OBB) and DOTA (HBB) are shown in Fig. 5(c) and (d), respectively. These results demonstrate the robustness of the detection process in our mSODANet and ensure its consistency on different types of objects under various scenarios. This is due to the fact that the employment of hierarchical dilated convolution network and the BFAM in conjunction with EfficientNet backbone effectively captures different contextual information of the objects that helps in improving the detection performance. It is also observed that the hierarchical dilated convolution network in mSODANet also performs equally well on arbitrary sized objects with different contextual information.

4.5. Ablation study

We explored various input image resolutions to feed into the backbone EfficientNet network to accommodate the various input resolutions of three benchmark object detection datasets. It is found from the experimental results that EfficientNet B3 backbone network works well on NWPU VHR-10 dataset in comparison to VisDrone2019 and DOTA datasets. This is due to the closeness of the input resolution of images in NWPU VHR-10 dataset (896) while the input resolution of VisDrone2019 is 1200 & DOTA is 1000. This is the reason to opt EfficientNet-B5 backbone network whose input resolution is 1280 for VisDrone2019 & DOTA datasets.

Table 5
Class-wise average precision (AP) comparison of mSODANet with state-of-the-art methods on DOTA OBB dataset.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
EfficientDet (B5) [32]	88.43	79.84	52.11	71.58	69.36	72.92	72.75	87.49	80.46	85.32	61.87	61.94	71.05	75.48	57.31
ROI Transformer [17]	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67
SCRDet [19]	89.98	80.65	52.09	75.92	68.81	73.68	83.59	90.85	87.94	86.86	65.02	66.68	67.00	68.24	65.21
GSDet [7]	81.12	76.78	40.78	75.89	64.50	58.37	74.21	89.92	79.40	78.83	64.54	63.67	66.04	58.01	52.13
HSP-Net [10]	90.42	86.91	62.57	79.96	78.13	81.86	**85.27**	90.80	87.30	85.94	69.96	72.11	**84.13**	80.99	**69.88**
mSODANet - E_6, E_7 (Ours)	**92.32**	**90.53**	**64.39**	**83.36**	**81.75**	**83.89**	83.11	**92.93**	**90.73**	**89.84**	**73.96**	**72.87**	83.90	**87.79**	68.62

Table 6
Class-wise average precision (AP) comparison of mSODANet with state-of-the-art methods on DOTA HBB dataset.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
CAD-Net [6]	87.80	82.40	49.40	73.50	71.10	63.50	76.70	90.90	79.20	73.30	48.40	60.90	62.00	67.00	62.20
EfficientDet (B5) [32]	89.23	82.52	54.90	73.90	71.45	75.38	75.69	89.92	81.56	87.73	63.05	63.56	74.19	78.86	59.13
SCRDet [19]	90.18	81.88	55.30	73.29	72.09	77.65	78.06	90.91	82.44	86.39	64.53	63.45	75.77	78.21	60.11
HSP-Net [10]	90.42	86.91	62.57	79.96	78.13	81.86	85.27	90.80	87.30	85.94	69.96	72.11	84.13	80.99	69.88
mSODANet - E_6, E_7 (Ours)	**95.58**	**92.31**	**68.42**	**85.98**	**87.12**	**87.88**	**90.17**	**95.21**	**94.63**	**90.26**	**74.89**	**75.92**	**88.74**	**84.83**	**76.07**

Note: The short names for categories are defined as PL-Plane, BD-Baseball diamond, BR-Bridge, GTF-Ground track field, SV-Small vehicle, LV-Large vehicle,SH-Ship, TC-Tennis court, BC-Basketball court, ST-Storage tank, SBF-Soccer-ball field, RA-Roundabout, HA-Harbor, SP-Swimming pool, and HC-Helicopter.

Table 7
Class-wise average precision (AP) comparison of mSODANet with state-of-the-art methods on NWPU VHR10 dataset.

Method	mAP (%)	Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle
EfficientDet (B3) [32]	81.51	95.09	68.83	74.69	81.54	72.05	83.27	79.87	85.55	91.07	83.20
Faster RCNN [2]	84.50	90.90	86.30	90.50	98.20	89.70	69.60	**100.0**	80.10	61.50	78.10
Li et al. [12]	87.10	99.70	90.80	90.60	92.90	90.30	80.10	90.80	80.30	68.50	87.10
CAD-Net [6]	91.50	97.00	77.90	95.60	93.60	87.60	87.10	99.60	**100.0**	86.20	89.90
R^2 CNN++ [44]	91.75	100.0	89.41	97.22	97.00	83.15	87.54	99.17	99.40	74.51	90.10
HSP-Net [10]	93.38	99.79	**92.45**	96.96	98.55	90.37	91.48	99.04	88.90	87.14	89.07
mSODANet - E_6, E_7 (Ours)	**97.81**	**100.0**	88.14	**99.46**	**100.0**	**98.92**	**99.11**	98.73	98.26	**97.65**	**98.23**

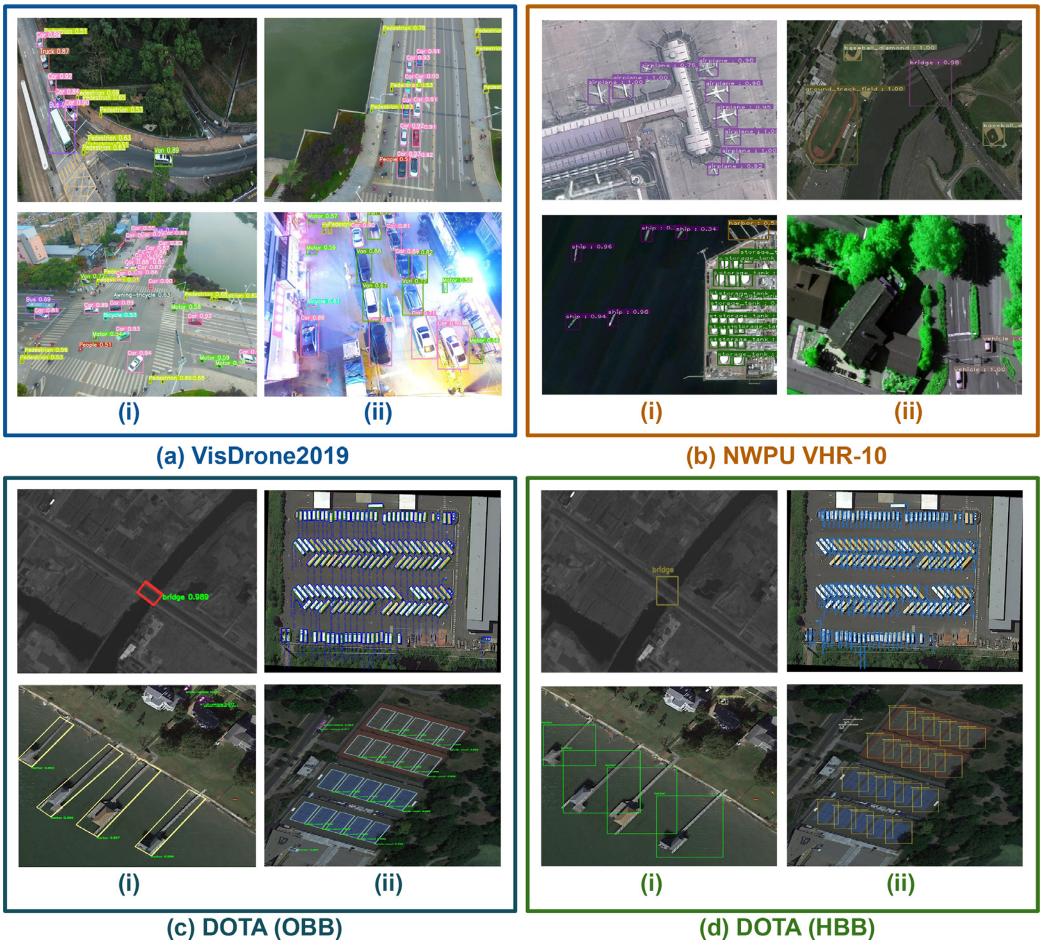


Fig. 5. The detection results on some images using mSODANet (our proposed method) with different characteristics. (a) VisDrone2019: (i) Arbitrary sized objects with multiple field-of-views (ii) Objects at different illuminations; (b) NWPU VHR-10: (i) Objects with different rotations and arbitrary spatial arrangements (ii) Different types of multi-scale objects; (c) DOTA (OBB) & (d) DOTA (HBB): (i) Multi-scale objects with arbitrary spatial distribution (ii) Densely-packed multi-scale objects.

Table 8
Threshold-wise average precision (AP) comparison of mSODANet on VisDrone2019 dataset for different dilation settings.

Method	AP (%)	AP_{50} (%)	AP_{75} (%)	AR_1 (%)	AR_{10} (%)	AR_{100} (%)	AR_{500} (%)
EfficientDet (B5) [32]	21.40	38.60	20.20	0.59	04.12	22.38	31.75
mSODANet - E_7 (Ours)	34.33	56.63	33.14	0.96	9.69	36.35	48.02
mSODANet - E_4, E_5 (Ours)	33.37	52.58	34.25	0.83	8.64	33.14	45.72
mSODANet - E_6, E_7 (Ours)	**36.89**	**55.92**	**37.41**	**1.15**	**11.36**	**37.25**	**48.92**

Table 9
Threshold-wise average precision (AP) comparison of mSODANet on DOTA (OBB) dataset for different dilation settings.

Method	AP (%)	AP_{50} (%)	AP_{75} (%)
EfficientDet (B5) [32]	72.52	75.86	73.36
mSODANet - E_7 (Ours)	83.41	87.23	81.76
mSODANet - E_4, E_5 (Ours)	79.34	83.92	78.68
mSODANet - E_6, E_7 (Ours)	**82.66**	**86.48**	**81.13**

Table 10
Threshold-wise average precision (AP) comparison of mSODANet on DOTA HBB dataset for different dilation settings.

Method	AP (%)	AP_{50} (%)	AP_{75} (%)
EfficientDet (B5) [32]	74.70	80.30	78.60
mSODANet - E_7 (Ours)	86.93	89.46	85.38
mSODANet - E_4, E_5 (Ours)	82.78	86.94	87.73
mSODANet - E_6, E_7 (Ours)	**85.83**	**90.33**	**89.61**

Table 11
Threshold-wise average precision (AP) comparison of mSODANet on NWPU-VHR10 dataset for different dilation settings.

Method	AP (%)	AP_{50} (%)	AP_{75} (%)
EfficientDet [32]	81.51	83.64	80.73
mSODANet - E_7 (Ours)	97.23	98.47	95.81
mSODANet - E_4, E_5 (Ours)	94.82	95.23	93.06
mSODANet - E_6, E_7 (Ours)	**97.81**	**99.15**	**96.33**

Table 12
Floating operations (FLOPs) comparison between proposed mSODANet and EfficientDet.

Method	Dataset	AP (%)	Parameters	GFLOPs
EfficientDet (B3)	NWPU VHR10	81.51	12 M	1.8
mSODANet (B3)	NWPU VHR10	**97.81**	13 M	1.93
EfficientDet (B5)	VisDrone2019	21.40	20 M	9.9
mSODANet (B5)	VisDrone2019	**36.89**	22 M	10.75

Further, we place DN_1 and DN_2 of hierarchical dilated network (HDN) at various levels of multi scale contextual features to investigate the effectiveness of the proposed mSODANET. The threshold-wise average precision of different HDN combinations on VisDrone 2019, DOTA OBB, DOTA HBB, and NWPU-VHR 10 can be seen fromTables 8–11.

From Tables 8–11, we can infer that the utilization of E_6 and E_7 features for DN_1 and DN_2 of HDN leads to state-of-the-art performance when compared to E_4 and E_5 for DN_1 and DN_2. Also, we can observe the performance deflation when using single DN_1 (E_7) component over two dilation components (DN_1 and DN_2). Moreover, the proposed mSODANet achieves significant improvement over EfficientDet [32]. This is due to the incorporation of contextual information of objects (small, medium, and large) at various scales.

Table 12 presents the model complexity of our proposed mSODANet in terms of floating point operations (FLOPs). It can be observed from the table that our proposed approach achieves better multi-scale object detection performance than EfficientDet [32] and uses slightly more FLOPs due to additional hierarchical dilation operations.

5. Conclusion

Due to the large variation in object sizes, distribution of dense objects, view-point, occlusions, and illumination changes, the object detection task in aerial images has become extremely challenging task. To address these challenges, in this paper, we propose a novel network, mSODANet for multi-scale object detection in aerial images. Specifically, we jointly combine the backbone network, hierarchical dilated network, and bi-directional feature aggregation module (BFAM) systematically to learn the efficient multi-scale contextual information of objects. The backbone network helps to extract robust multi-scale features. And, the hierarchical dilated network learns the contextual information of objects at multiple-scales and multiple field-of-views. The obtained multi-scale features are further refined with BFAM to incorporate effective representation of varied size and dense objects. Thus, the proposed object detection framework learns the characteristics of different objects in aerial images and further enhances multi-scale object detection capability in the network. The performance of the proposed mSODANet is evaluated on three challenging aerial image datasets and reported state-of-the art performance on all three datasets. Moreover, the quantitative and qualitative analysis present the effectiveness of proposed approach and demonstrate the robustness of the model. However, our proposed approach requires large amount of storage capability for effective data pooling to a centralized location. To overcome such limitations, we can explore federated learning for computation-effective training in the future. Also, we can extend mSODANet to preserve data privacy in object detection tasks in aerial images through adversarial attacks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 21–37.

[2] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.

[3] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303338.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.

[5] G. Cheng, P. Zhou, J. Han, RIFD-CNN: rotation-invariant and fisher discriminative convolutional neural networks for object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2884–2893.

[6] G. Zhang, S. Lu, W. Zhang, CAD-Net: a context-aware detection network for objects in remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 57 (12) (2019) 10015–10024.

[7] W. Li, W. Wei, L. Zhang, GSDet: object detection in aerial images based on scale reasoning, IEEE Trans. Image Process. 30 (2021) 4599–4609.

[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[10] C. Xu, C. Li, Z. Cui, T. Zhang, J. Yang, Hierarchical semantic propagation for object detection in remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 58 (6) (2020) 4353–4364.

[11] G. Cheng, P. Zhou, J. Han, Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 54 (12) (2016) 7405–7415.

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

[13] W. Guo, W. Yang, H. Zhang, G. Hua, Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network, Remote Sens. 10 (1) (2018) 131.

[14] B. Singh, M. Najibi, L.S. Davis, SNIPER: efficient multi-scale training, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, 2018.

[15] M. Najibi, B. Singh, L. Davis, Autofocus: efficient multi-scale inference, 2019,

[16] M. Mandal, M. Shah, P. Meena, S. Devi, S.K. Vipparthi, AVDNet: a small-sized vehicle detection network for aerial visual data, IEEE Geosci. Remote Sens. Lett. 17 (3) (2019) 494–498.

[17] J. Ding, N. Xue, Y. Long, G.-S. Xia, Q. Lu, Learning ROI transformer for oriented object detection in aerial images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2849–2858.

[18] F. Yang, H. Fan, P. Chu, E. Blasch, H. Ling, Clustered object detection in aerial images, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8310–8319.

[19] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, K. Fu, SCRDet: towards more robust detection for small, cluttered and rotated objects, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8232–8241.

[20] Z. Zou, Z. Shi, Random access memories: a new paradigm for target detection in high resolution aerial remote sensing images, IEEE Trans. Image Process. 27 (2018) 1100–1111.

[21] C. Li, T. Yang, S. Zhu, C. Chen, S. Guan, Density map guided object detection in aerial images, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 737–746.

[22] W. Yu, T. Yang, C. Chen, Towards resolving the challenge of long-tail distribution in UAV images for object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3258–3267.

[23] B. Bosquet, M. Mucientes, V.M. Brea, STDnet-ST: spatio-temporal ConvNet for small object detection, Pattern Recognit. 116 (2021) 107929.

[24] H. Wang, Q. Wang, P. Li, W. Zuo, Multi-scale structural kernel representation for object detection, Pattern Recognit. 110 (2021) 107593.

[25] W. Ma, Y. Wu, F. Cen, G. Wang, MDFN: multi-scale deep feature learning network for object detection, Pattern Recognit. 100 (2020) 107–149.

[26] Y. Kong, M. Feng, X. Li, H. Lu, X. Liu, B. Yin, Spatial context-aware network for salient object detection, Pattern Recognit. 114 (2021) 107867.

[27] Q. Zhang, Y. Shi, X. Zhang, Attention and boundary guided salient object detection, Pattern Recognit. 107 (2020) 107484.

[28] S. Mohammadi, M. Noori, A. Bahri, S.G. Majelan, M. Havaei, CAGNet: content-aware guidance for salient object detection, Pattern Recognit. 103 (2020) 107303.

[29] K. Shuang, Z. Lyu, J. Loo, W. Zhang, Scale-balanced loss for object detection, Pattern Recognit. 117 (2021) 107997.

[30] Z. Huang, H.-X. Chen, T. Zhou, Y.-Z. Yang, C.-Y. Wang, B.-Y. Liu, Contrast-weighted dictionary learning based saliency detection for VHR optical remote sensing images, Pattern Recognit. 113 (2021) 107757.

[31] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, ICML, 2019, pp. 6105–6114.

[32] M. Tan, R. Pang, Q.V. Le, EfficientDet: scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[33] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.

[34] Q. Yao, X. Hu, H. Lei, Geospatial object detection in remote sensing images based on multi-scale convolutional neural networks, in: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 1450–1453.

[35] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.

[36] G. Ghiasi, T.-Y. Lin, Q.V. Le, NAS-FPN: learning scalable feature pyramid architecture for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7036–7045.

[37] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, H. Ling, Vision meets drones: past, present and future.(2020). arXiv:2001.06303

[38] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, DOTA: alarge-scale dataset for object detection in aerial images, 2019, arXiv:1711.10398

[39] G. Cheng, J. Han, P. Zhou, L. Guo, Multi-class geospatial object detection and geographic image classification based on collection of part detectors, ISPRS J. Photogramm. Remote Sens. 98 (2014) 119–132.

[40] H. Wang, Z. Wang, M. Jia, A. Li, T. Feng, W. Zhang, L. Jiao, Spatial attention for multi-scale feature refinement for object detection, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 64–72.

[41] C. Chen, Y. Zhang, Q. Lv, S. Wei, X. Wang, X. Sun, J. Dong, RRNet: a hybrid detector for object detection in drone-captured images, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 100–108.

[42] S. Hong, S. Kang, D. Cho, Patch-level augmentation for object detection in aerial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019. 0–0

[43] S. Zhang, G. He, H.-B. Chen, N. Jing, Q. Wang, Scale adaptive proposal network for object detection in remote sensing images, IEEE Geosci. Remote Sens. Lett. 16 (6) (2019) 864–868.

[44] J. Pang, C. Li, J. Shi, Z. Xu, H. Feng, R^2-CNN: fast tiny object detection in large-scale remote sensing images, IEEE Trans. Geosci. Remote Sens. 57 (8) (2019) 5512–5524.

Chalavadi Vishnu received his B.Tech (CSE) from JNTU Hyderabad in 2016 and M.Tech (CSE) from IIT Hyderabad, in 2018, in India. Currently he is pursuing PhD (CSE) at IIT Hyderabad. His research interests include learning representations of video activities and drone data, and autonomous vehicles.

Jeripothula Prudviraj received B.Tech degree from Sreenidhi Institute of Science and Technology, Hyderabad, India, in 2013, the M.Tech degree from MANIT Bhopal, India, in 2015. pursuing PhD (CSE) at IIT Hyderabad (India). His research interests include video content analysis, deep learning, and computer vision.

Rajeshreddy Datla received B.Tech. (2002) and M.Tech. (2012) in CSE from JNTU Hyderabad, India. He has completed his PhD (CSE) at IIT Hyderabad (India) in 2021. He is currently working as Scientist/Engineer-SE in ADRIN, Department of Space, Secunderabad, India. His research interests include remote sensing imagery analysis, machine learning, and deep learning.

Dr. Sobhan Babu received B.E from University of Madras, in 1999, M.Tech (CSE) from IIT Bombay, in 2001 and PhD (CSE) from Institute of Technology Bombay (IITB), in 2007, in India. He is currently an Associate Professor in CSE at IIT Hyderabad, India. His research interests include big data analytics, graph theory and algorithms.

Dr. C. Krishna Mohan received M.Tech (SACA) from NIT Surathkal in 2000, and PhD (CSE) from IIT Madras in 2007, in India. He is currently Professor in the dept. of CSE, IIT Hyderabad (India). His research interests include video content analysis and machine learning. He is a senior member of IEEE.