

OA-CapsNet: A One-Stage Anchor-Free Capsule Network for Geospatial Object Detection from Remote Sensing Imagery

Yongtao Yu, Junyong Gao, Chao Liu, Haiyan Guan, Dilong Li, Changhui Yu, Shenghua Jin, Fenfen Li & Jonathan Li

To cite this article: Yongtao Yu, Junyong Gao, Chao Liu, Haiyan Guan, Dilong Li, Changhui Yu, Shenghua Jin, Fenfen Li & Jonathan Li (2021): OA-CapsNet: A One-Stage Anchor-Free Capsule Network for Geospatial Object Detection from Remote Sensing Imagery , Canadian Journal of Remote Sensing, DOI: [10.1080/07038992.2021.1898937](https://doi.org/10.1080/07038992.2021.1898937)

To link to this article: <https://doi.org/10.1080/07038992.2021.1898937>



Published online: 17 Mar 2021.



Submit your article to this journal [↗](#)



Article views: 18







View related articles [↗](#)



View Crossmark data [↗](#)

OA-CapsNet: A One-Stage Anchor-Free Capsule Network for Geospatial Object Detection from Remote Sensing Imagery

OA-CapsNet: Un réseau capsule sans ancrage en une seule étape pour la détection d'objets géospatiaux à partir d'images de télédétection

Yongtao Yu^a , Junyong Gao^a, Chao Liu^a, Haiyan Guan^b , Dilong Li^c , Changhui Yu^a, Shenghua Jin^a, Fenfen Li^a, and Jonathan Li^d 

^aFaculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, China; ^bSchool of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China; ^cState Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China; ^dDepartment of Geography and Environmental Management, University of Waterloo, Waterloo, Canada

ABSTRACT

Object detection from remote sensing images serves as an important prerequisite to many applications. However, caused by scale and orientation variations, appearance and distribution diversities, occlusion and shadow contaminations, and complex environmental scenarios of the objects in remote sensing images, it brings great challenges to realize highly accurate recognition of geospatial objects. This paper proposes a novel one-stage anchor-free capsule network (OA-CapsNet) for detecting geospatial objects from remote sensing images. By employing a capsule feature pyramid network architecture as the backbone, a pyramid of high-quality, semantically strong feature representations are generated at multiple scales for object detection. Integrated with two types of capsule feature attention modules, the feature quality is further enhanced by emphasizing channel-wise informative features and class-specific spatial features. By designing a centreless-assisted one-stage anchor-free object detection strategy, the proposed OA-CapsNet performs effectively in recognizing arbitrarily-orientated and diverse-scale geospatial objects. Quantitative evaluations on two large remote sensing datasets show that a competitive overall accuracy with a precision, a recall, and an F_{score} of 0.9625, 0.9228, and 0.9423, respectively, is achieved. Comparative studies also confirm the feasibility and superiority of the proposed OA-CapsNet in geospatial object detection tasks.

RÉSUMÉ

La détection d'objets à partir d'images de télédétection constitue une condition préalable importante à de nombreuses applications. Cependant, les variations d'échelle et d'orientation, les diversités d'apparence et de distribution, les contaminations de l'occlusion et de l'ombre, et des scénarios environnementaux complexes des objets dans les images apportent de grands défis pour réaliser la reconnaissance très précise des objets géospatiaux. Cet article propose un nouveau réseau capsule sans ancrage en une seule étape (OA-CapsNet) pour détecter les objets géospatiaux à partir d'images de télédétection. En utilisant une architecture de réseau pyramidal à capsules comme colonne vertébrale, une pyramide de représentations de fonctionnalités de haute qualité et sémantiquement fortes sont générées à plusieurs échelles pour la détection d'objets. Intégrée à deux types de modules à capsules, la qualité des fonctionnalités est encore améliorée en mettant l'accent sur les fonctionnalités informatives du côté des canaux et des caractéristiques spatiales spécifiques à la classe. En concevant une stratégie de détection d'objets sans ancrage à un étage assistée par la centralité, l'OA-CapsNet proposé fonctionne efficacement dans la reconnaissance d'objets géospatiaux arbitrairement orientés et diversifiés. Les évaluations quantitatives sur deux grands ensembles de données de télédétection montrent qu'une exactitude concurrentielle globale est atteinte avec une précision, un *recall* et un *Fscore* de 0,9625, 0,9228 et 0,9423 respectivement. Des études comparatives confirment également la faisabilité et la supériorité de l'OA-CapsNet dans les tâches de détection d'objets géospatiaux.

ARTICLE HISTORY

Received 19 November 2020
Accepted 28 February 2021

Introduction

With the increasing advancement of optical remote sensing sensors in flexibility, cost-efficiency, resolution, and quality, remote sensing images have become an important data source for many applications, such as land cover mapping, landmark recognition, environmental analysis, and intelligent transportation systems. As a typical research topic, geospatial object detection targets to correctly identify the objects of interest and accurately locate their positions in the remote sensing image. To date, a number of techniques with continuously improving performances have been developed for geospatial object detection tasks in the literature. However, it is still a challenging issue to fulfill highly accurate and fully automated detection of geospatial objects due to the complicated scenarios of the geospatial objects in the bird-view remote sensing images, such as scale and orientation variations, texture and distribution diversities, occlusion and shadow contaminations, illumination condition changes, and complex environmental conditions. Thus, exploiting advanced techniques to further enhance the accuracy and automation level of geospatial object detection is greatly meaningful and positively favorable to a wide range of applications.

Recent increasing development of deep learning techniques has attracted great attentions on geospatial object detection by using deep learning models (Li, Wan, et al. 2020). Specifically, the existing deep learning based geospatial object detection methods can be roughly categorized into one-stage methods and two-stage methods. The two-stage methods comprise two cascaded processing modules for, respectively, region proposal generation and object recognition. Usually, a region proposal subnetwork is designed to generate a set of dense object region proposals, which are further verified by a classification subnetwork for recognizing the objects of interest (Gong et al. 2020; Li et al. 2019; Liu et al. 2021; Yu et al. 2020; Zheng et al. 2020). In contrast, the one-stage methods accomplish feature extraction and object detection with a single network without the pre-generation of object region proposals. Hu et al. (2019) proposed a convolutional neural network (CNN), which was trained with a sample updating strategy, to detect objects in large-area remote sensing images. The updated artificial composite samples were used to fine-tune the object detector. Based on the YOLOv2 architecture, Liu et al. (2019) designed a multilayer feature concatenation and feature introducing strategy aiming to improve the adaptability of the network to multiscale objects, especially the small-size objects. By extracting multiscale

features and learning visual attentions at each feature scale, Wang et al. (2019) developed a multiscale visual attention network (MS-VAN) to detect multiscale objects. The multiscale features were extracted and properly fused through a skip-connected encoder-decoder backbone. To upgrade the detection accuracy of small-size objects, Qin et al. (2021) proposed a specially optimized one-stage network (SOON), which comprised three parts of feature enhancement, multi-scale detection, and feature fusion. In this network, the spatial information of small-size objects was concentrated on by incorporating a receptive field enhancement module. Similarly, a one-stage network integrated with residual blocks at multiple scales was constructed by Mandal et al. (2020) for detecting small-size vehicles. The residual blocks, alongside the enlarged output feature map, enhanced the representation robustness of the feature saliencies for small-size objects. Tang et al. (2017) proposed an orientated single-shot multi-box detector (SSD) for detecting arbitrarily-orientated vehicles. This model deployed a set of default orientated anchors with varying scales at each position of the feature map to produce detection bounding boxes. Zhang, Liu, et al. (2020) designed a depthwise-separable attention-guided network (DAGN) to detect vehicles by integrating a feature concatenation and attention block into the YOLOv3 architecture. With the combination of the multi-level feature concatenation and channel feature attention mechanisms, the feature representation capability of the network was dramatically enhanced to serve for the small-size vehicles. In addition, YOLO-fine (Pham et al. 2020) and H-YOLO (Tang et al. 2020) models were also developed based on the YOLOv3 architecture to detect geospatial objects. Specifically, the YOLO-fine model performed successfully in detecting small-size objects with both high accuracy and high speed, which was applicable for real-time applications; whereas, the H-YOLO model combined the region of interest preselected network and textural properties to effectively improve the object detection accuracy.

Yao et al. (2021) proposed a multiscale CNN (MSCNN) based on an EssNet backbone and a dilated bottleneck block for extracting multiscale high-quality features. The EssNet backbone functioned to maintain the resolution of deep feature levels and improve the feature encoding of the multiscale objects. To effectively handle arbitrarily-orientated objects, Zhou, Zhang, Gao, et al. (2020a) developed a rotated feature network (RFN), which generated rotation-aware features to delineate orientated objects and rotation-invariant features to conduct object recognition.

Bao et al. (2019) designed a single-shot anchor refinement network (S²ARN) for detecting orientated objects with the assistance of orientated bounding boxes. In this model, two types of regressions were applied to, respectively, generate refined anchors and accurate bounding boxes. Differently, Shi et al. (2020) constructed a one-stage anchor-free network to detect arbitrarily-orientated vehicles. In this network, a feature pyramid fusion strategy was applied to concatenate the multi-stage features for the direct regression and identification of vehicles. Similarly, Zhang, Wang, et al. (2020) proposed an anchor-free network for detecting rotated ships. This network consisted of a feature extraction backbone integrated with a selective concatenation module, a rotation Gaussian-mask module for modeling the geometric features of ships, and a detection module for ship detection and rotated bounding box regression. Li, Pei, et al. (2020) also presented a single-stage anchor-free detector based on a multiscale dense path aggregation feature pyramid network (DPAFPN). The DPAFPN performed promisingly in comprehensively considering high-level semantic information and low-level location information and avoiding information loss during shallow feature transfer. Wu et al. (2018) learned a regularized CNN to extract multiscale and rotation-insensitive convolutional channel features. These features were finally fed into an outlier removal assisted AdaBoost classifier for object recognition. To effectively tackle small-size objects, Courtrai et al. (2020) designed a generative adversarial network (GAN). Specifically, a super-resolution technique with an object-focused strategy was applied to highlight the details of the small-size objects. Likewise, by combining super-resolution and edge enhancement techniques, Rabbi et al. (2020) proposed an edge-enhanced super-resolution GAN (EESRGAN) and used varying detector networks in an end-to-end manner for object detection. Mekhalfi et al. (2019) designed a capsule network for detecting objects in unmanned aerial vehicle (UAV) images. The capsule network comprised a conventional convolutional layer and a capsule convolutional layer for capsule feature extraction, and a fully-connected capsule layer for object recognition. In addition, Siamese graph embedding network (SGEN) (Tian et al. 2020), feature-merged single-shot detection network (FMSSD) (Wang et al. 2020), single-shot multiscale feature fusion network (Zhuang et al. 2019), single-shot recurrent network with activated semantics (Chen et al. 2018), fully convolutional network (FCN) (Cozzolino et al. 2017), region-enhanced CNN (RECNN) (Lei et al. 2020), and

Bayesian transfer learning (Zhou, Zhang, Liu et al. 2020b) were also leveraged to detect geospatial objects.

In this paper, we develop a novel one-stage anchor-free capsule network for detecting geospatial objects from remote sensing images. With the formulation of a capsule feature pyramid network architecture as the backbone for extracting multiscale high-order capsule features, the integration of the capsule-based channel and spatial feature attention modules for obtaining informative feature encodings, and the design of an effective anchor-free object detection strategy, the proposed network performs promisingly in handling geospatial objects of different scales, orientations, distributions, and surface conditions in diverse complicated scenarios. The contributions of this paper include the following: (1) two types of capsule feature attention modules are proposed to emphasize channel-wise informative features and class-specific spatial features to produce high-quality object-orientated feature representations; (2) an effective centreness-assisted one-stage anchor-free object detection strategy is designed to recognize arbitrarily-orientated and varying-scale geospatial objects.

Methodology

Capsule network

Traditional deep learning models are generally designed with scalar neurons for encoding the feature saliencies and probabilities. In contrast, constructed with vectorial capsules, capsule networks use the capsule length to encode the feature probability of an entity and the instantiation parameters of a capsule to depict the inherent features of the entity (Sabour et al. 2017). An advantageous property of the capsule formulation is that the vectorial representation allows a capsule not only to detect a feature but also to learn and identify its variants. That is, a category of entity features can be encoded by using capsules rather than a single feature encoding pattern like that in the CNN. Capsule convolutions operate quite differently from traditional convolutions. Specifically, for a capsule j , the input to the capsule is a weighted aggregation over all the predictions from the capsules within the convolution kernel in the previous layer as follows:

$$C_j = \sum_i a_{i,j} \times U_{i,j} \quad (1)$$

where C_j is the aggregated input to capsule j ; $a_{i,j}$ is a coupling coefficient reflecting the contribution of capsule i to capsule j , which is dynamically determined

by the improved dynamic routing process (Rajasegaran et al. 2019); $U_{i,j}$ is the prediction from capsule i to capsule j , which is computed as follows:

$$U_{i,j} = \mathbf{W}_{i,j}U_i \quad (2)$$

where U_i is the output of capsule i and $\mathbf{W}_{i,j}$ is a transformation matrix acting as a feature mapping function.

As for the capsule length-based feature probability encoding pattern in the capsule networks, the longer the capsules are, the higher the probability predictions should be. To this end, a squashing function (Sabour et al. 2017) is designed as the activation function to normalize the output of a capsule. The squashing function is formulated as follows:

$$U_j = \frac{\|C_j\|^2}{1 + \|C_j\|^2} \times \frac{C_j}{\|C_j\|} \quad (3)$$

where C_j and U_j are, respectively, the input and the output of capsule j . The modulus of a vector is calculated by the operator $\|\cdot\|$ for representing the capsule length. Through the above normalization process, long capsules are shrunk to a length close to one to cast high predictions; whereas short capsules are suppressed to almost a zero length to provide few contributions.

One-stage anchor-free capsule network

As shown in Figure 1, based on capsule representations, we construct a fully convolutional one-stage anchor-free capsule network (OA-CapsNet) for detecting geospatial objects. The architecture of the OA-CapsNet involves a feature extraction backbone and a set of parallel multiscale object detection heads. By employing a feature pyramid network architecture, the feature extraction backbone serves to extract high-quality, multiscale capsule features. The object detection head is designed with a one-stage anchor-free strategy for directly detecting and regressing objects, which avoids the tedious work in anchor determination and proposal generation, as well as improving the detection efficiency.

The feature extraction backbone comprises two traditional convolutional layers for extracting low-level image features, and a pyramid of capsule convolutional layers for extracting multiscale high-level entity features. For the traditional convolutional layers, the rectified linear unit (ReLU) is adopted as the activation function. The scalar features output by the second traditional convolutional layer are transformed into vectorial capsule representations to constitute the primary capsule layer for further characterizing entity

features. This can be achieved through traditional convolution operations, followed by feature channel grouping and capsule vectoring. The capsule convolutional layers are split into four network stages by three capsule max-pooling layers to extract capsule features at different scales with a scaling step of two. Specifically, within each stage, the feature maps maintain the same spatial resolution and size. The spatial size of feature maps is gradually scaled down stage by stage to produce lower-resolution, but semantically higher-level, feature maps. For each stage, the feature map of the top layer, which encodes the strongest and the most representative feature semantics, is selected and modulated with a 1×1 capsule convolution to form a reference feature map for further feature fusion and augmentation. As shown in Figure 1, the set of reference feature maps obtained from the four stages is denoted by $\{G_1, G_2, G_3, G_4\}$. Then, through a series of operations, including capsule deconvolutions to upsample a high-level reference feature map to its twice spatial size, capsule feature concatenations to concatenate the upsampled high-level reference feature map with a low-level reference feature map from the previous stage, and capsule convolutions to conduct feature fusion, these multiscale reference feature maps are gradually fused in a top-down manner to generate a set of multiscale fused feature maps $\{D_1, D_2, D_3\}$. For instance, feature map G_4 is first upsampled to its twice spatial size to hallucinate a higher-resolution feature map through capsule deconvolution operations. Then, the upsampled feature map is concatenated with feature map G_3 through the lateral connection. Afterward, a 3×3 capsule convolution is performed on the concatenated feature maps to conduct feature fusion, resulting in the multiscale fused feature map D_3 . This process repeats downward to gradually fuse all the feature maps $\{G_1, G_2, G_3, G_4\}$. Finally, a 3×3 capsule convolution is applied to $\{D_1, D_2, D_3, G_4\}$ to further smooth the fused features to obtain the multiscale feature maps $\{P_1, P_2, P_3, P_4\}$, which have scales of $\{1, 1/2, 1/4, 1/8\}$ with regard to the input image and encode strong feature semantics at each scale. In this way, the high-resolution features in the lower stages are effectively augmented by the semantically strong features in the higher stages to provide a set of high-quality feature maps at multiple scales.

Specifically, to enhance the feature representation capability at each scale, we design a capsule-based channel feature attention (CFA) module and integrate it at the end of each stage to boost the quality of the reference feature map. The architecture of the CFA module is inspired by the squeeze-and-excitation

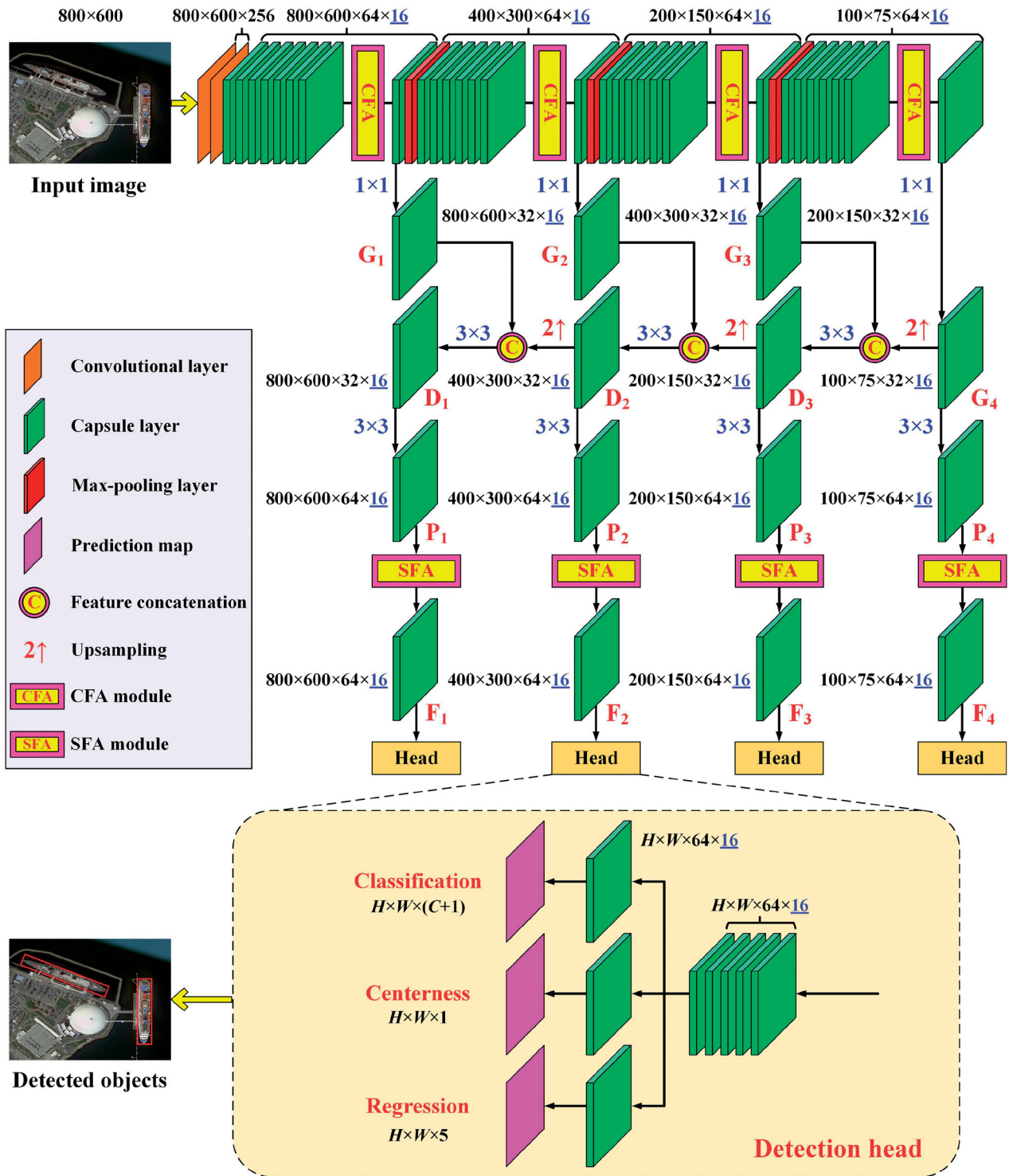


Figure 1. Architecture of the proposed one-stage anchor-free capsule network (OA-CapsNet). The dimension of a capsule is configured as 16.

network (Hu et al. 2018) and leverages a capsule-based formulation. The CFA module aims to upgrade the input features by exploiting the channel-wise interdependencies to increase the ability of the network to highlight informative feature channels associated with the foreground and suppress the impacts of the helpless or less informative feature channels. To

this end, as shown in Figure 2, first, the input multi-dimensional capsule feature map is converted into a one-dimensional capsule feature map A , which maintains the same number of feature channels, as well as the same spatial size, and mainly encodes the feature saliencies of the input feature map, through a 1×1 capsule convolution. Then, a channel descriptor C is

obtained by performing a global average pooling on feature map A in a channel-wise manner. That is, a scalar value reflecting the global feature statistics in a feature channel is generated by spatially averaging the features in this channel. Finally, two fully-connected layers are connected to exploit channel-wise interdependencies. These two fully-connected layers are, respectively, activated by the ReLU and the sigmoid function. The output of the second fully-connected layer produces a channel-wise attention descriptor R , whose length equals to the number of channels of the input feature map and each of whose entries encodes the importance of the associated channel of the input feature map. That is, the larger the value of an entry, the more important the associated channel of the input feature map. Whereas, the smaller the value of an entry, the less informative the associated channel of the input feature map. The attention descriptor R is used as weight factors to recalibrate the input feature map to emphasize the informative and salient features and weaken the contributions of the less important ones. This is achieved by multiplying R with the input feature map in a channel-wise and element-wise manner as follows:

$$\bar{U}_j^i = r_i \times U_j^i, i = 1, 2, \dots, 64 \quad (4)$$

where r_i represents the i -th entry of the attention

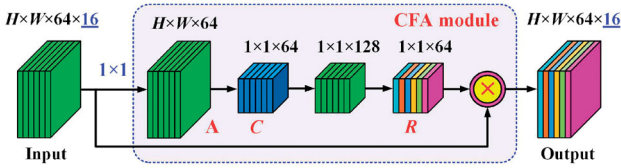


Figure 2. Architecture of the channel feature attention (CFA) module.

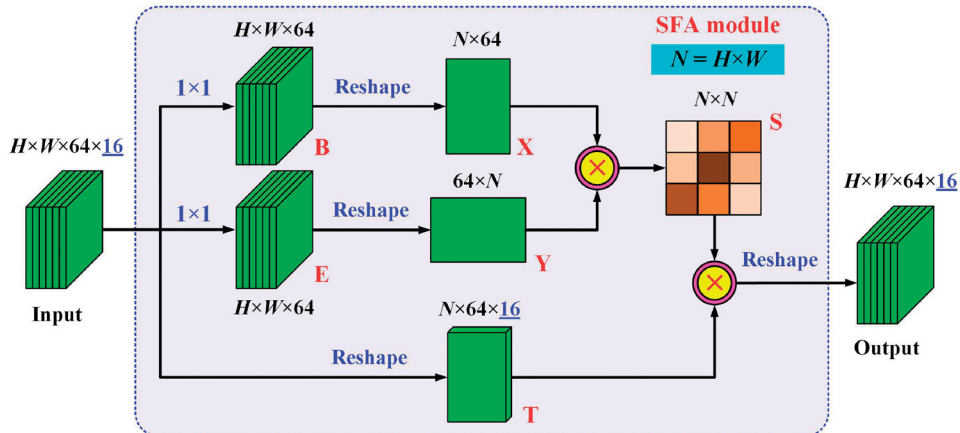


Figure 3. Architecture of the spatial feature attention (SFA) module.

descriptor R ; U_j^i and \bar{U}_j^i are, respectively, the output of the original capsule and the output of the recalibrated capsule in the i -th channel. In this way, the informative feature channels are significantly highlighted and the less informative feature channels are rationally suppressed, thereby boosting the feature representation robustness.

In addition, to further enhance the feature semantics at each scale, we design a capsule-based spatial feature attention (SFA) module and mount it over the multiscale feature maps $\{P_1, P_2, P_3, P_4\}$. The architecture of the SFA module is inspired by the dual attention network (Fu et al. 2019) and adopts a capsule-based formulation. The SFA module functions to enforce the network to focus on the spatial features associated with the object regions and adequately suppress the influences of the background features. As shown in Figure 3, first, two 1×1 capsule convolutions are, respectively, performed on the input multi-dimensional capsule feature map to convert it into two one-dimensional identical-size feature maps $B \in \mathbb{R}^{H \times W \times 64}$ and $E \in \mathbb{R}^{H \times W \times 64}$, where H and W are, respectively, the height and width of the input feature map. Then, feature maps B and E are reshaped along different dimensions to obtain two feature matrices $X \in \mathbb{R}^{N \times 64}$ and $Y \in \mathbb{R}^{64 \times N}$, where $N = W \times H$ denotes the number of positions in the input feature map. Next, we multiply X with Y (i.e. XY) and activate each element of the product matrix by a softmax function in a column manner to constitute a spatial attention matrix $S \in \mathbb{R}^{N \times N}$. The element $s_{i,j}$ at row i , column j of the spatial attention matrix S encodes the impact of position i on position j in the input feature map. The computation of the element $s_{i,j}$ is as follows:

$$s_{i,j} = \frac{\exp(\sum_{k=1}^{64} x_{i,k} \times y_{k,j})}{\sum_{m=1}^N \exp(\sum_{k=1}^{64} x_{m,k} \times y_{k,j})} \quad (5)$$

where $x_{i,k}$ denotes the element at row i , column k of the feature matrix \mathbf{X} and $y_{k,j}$ denotes the element at row k , column j of the feature matrix \mathbf{Y} . Finally, the input feature map is reshaped to a capsule feature matrix $\mathbf{T} \in \mathbb{R}^{N \times 64 \times 16}$ and multiplied with the spatial attention matrix \mathbf{S} (i.e. \mathbf{ST}), followed by a reshaping operation to reshape the product matrix to the dimension $\mathbb{R}^{H \times W \times 64 \times 16}$, to produce a high-quality spatial feature highlighted feature map. In this way, the spatial features associated with the object regions are positively concentrated on and the background features are effectively weakened, thereby improving the object-level feature characterization quality. As shown in Figure 1, the multiscale class-specific feature maps $\{F_1, F_2, F_3, F_4\}$ output by the SFA modules are fed into the object detection head to conduct object inference. This set of feature maps comprehensively take into account both the channel and spatial feature informativeness, as well as the multiscale feature semantics, to provide semantically strong feature representations at multiple scales.

As shown in Figure 1, the object detection head employs a shallow capsule convolutional network with three parallel output branches for, respectively, identifying the presence of an object, inferring the centreness of a position, and regressing the orientated bounding box of an object. The classification branch outputs a $(C+1)$ -dimensional softmax vector at each position for recognizing the C categories of objects and the background. That is, the outputs at each position of the classification branch are activated by the softmax function to generate a one-hot prediction. As a result, the category corresponding to the maximum softmax output is assigned as the predicted category label at a position. Specifically, to effectively handle arbitrarily-orientated objects, we use a five-tuple representation $\{d_1, d_2, d_3, d_4, \theta\}$ to characterize an object at a position. As shown in Figure 4a, $\{d_1, d_2, d_3, d_4\}$ represent the distances from a position inside the object region to the four sides of the object's bounding box, and $\theta \in [0, \pi)$ represents the orientation of the

object, which is defined as the included angle from the positive direction of the x -axis to the direction parallel to the long side of the object's bounding box along the anticlockwise direction. Based on such representation, the regression branch outputs a five-dimensional vector at each position for encoding the orientated bounding box of an object. Note that, since the object scales vary greatly in the multiscale feature maps, thus, rather than directly regressing the large-range parameters $\{d_1, d_2, d_3, d_4\}$, we adopt the following transformations to restrain them to a small range:

$$d_i = \frac{\sqrt{W_F^2 + H_F^2}}{2} e^{b_i}, \quad i = 1, 2, 3, 4 \quad (6)$$

where W_F and H_F are, respectively, the width and height of a feature map used for regression. As a result, the regression branch only requires to output the small-range values $b_i, i = 1, 2, 3, 4$ in the feature map domain. The predicted regression parameters in the image domain can be computed using Equation (6). In addition, to effectively focus on the high-quality bounding boxes near the object center and suppress the low-quality bounding boxes at the positions far away from the object center, we add an indicator to depict the centreness of a position, which measures the proximity of a position to the object center. That is, the higher the centreness measure at a position, the closer the position to the object center. Thus, the centreness branch outputs a scalar value (activated by the sigmoid function) at each position for inferring the centreness of the position, which is used to weight the corresponding objectness score of the classification branch. Concretely, the certainty of the existence of an object at a position is reflected by the product of the softmax output from the classification branch and the centreness indicator from the centreness branch.

To construct a high-quality object detection model, at the training stage, the positive samples are selected as the positions located at the central area surrounded by the object's bounding box with a ratio of $\delta = 0.8$ (Figure 4b). The remaining marginal area containing

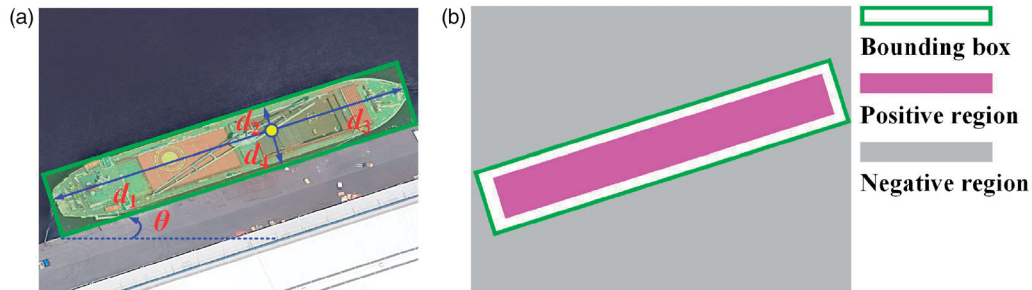


Figure 4. Illustrations of (a) the five-tuple representation of an arbitrarily-orientated object at a position, and (b) the positive and negative regions used for training.

low-quality object information is directly ignored. The background area is used as the negative samples. Based on the multi-branch prediction architecture of the OA-CapsNet, the loss function used for directing the training process is designed as a multitask loss function as follows:

$$L = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{reg}} + \lambda_3 L_{\text{cnt}} \quad (7)$$

where L_{cls} , L_{reg} , and L_{cnt} are, respectively, the classification, regression, and centreness loss terms of the three prediction branches; λ_1 , λ_2 , and λ_3 are the regularization factors for balancing the contributions of the three loss terms. To effectively handle orientated bounding boxes, the L_{reg} is formulated as the generalized intersection over union (GIoU) loss (Rezatofighi et al. 2019) between the regressed bounding boxes and the target bounding boxes. The L_{cls} and L_{cnt} are formulated as the focal loss (Lin et al. 2017) between the predictions and the ground truths. Specifically, given the distance regression targets $\{d_1^*, d_2^*, d_3^*, d_4^*\}$ for a position, the corresponding centreness target is calculated accordingly as follows:

$$c^* = \sqrt{\frac{\min(d_1^*, d_3^*)}{\max(d_1^*, d_3^*)} \times \frac{\min(d_2^*, d_4^*)}{\max(d_2^*, d_4^*)}} \quad (8)$$

Results and discussion

Datasets

We evaluated the geospatial object detection performance of the proposed OA-CapsNet on the following two large-scale remote sensing image datasets: GOD²18 (Yu et al. 2020) and DOTA (Xia et al. 2018) datasets. The GOD²18 dataset contains 22,000 images covering four categories of geospatial objects, including airplane, ship, vehicle, and ground track field. This dataset involves 69,207 annotated instances labeled with orientated bounding boxes. All the images have the same image size of 800×600 pixels. The DOTA dataset comprises 2086 images covering fifteen categories of geospatial objects, including vehicle, airplane, ship, harbor, bridge, tennis court, etc. This dataset involves 188,282 annotated instances labeled with arbitrary quadrilaterals. The image sizes range from about 800×800 pixels to 4000×4000 pixels. These two datasets are remarkably challenging, since the images were captured using different sensors and platforms and the geospatial objects exhibit with varying scales and orientations, diverse appearances and distributions, different-level occlusion and shadow contaminations, and complicated environmental scenarios. At the training stage, the training sets of these

two datasets were applied to construct the proposed OA-CapsNet.

Network training

The proposed OA-CapsNet was trained in an end-to-end manner by backpropagation and stochastic gradient descent on a cloud computing platform with ten 16-GB GPU, one 16-core CPU, and a memory size of 64 GB. Before training, we randomly initialized all layers of the OA-CapsNet by drawing parameters from a zero-mean Gaussian distribution with a standard deviation of 0.01. Each training batch contained two images per GPU and was trained for 1000 epochs. During training, we configured the initial learning rate as 0.001 for the first 800 epochs and decreased it to 0.0001 for the rest 200 epochs. The momentum and weight decay were configured as 0.9 and 0.0005, respectively. To trade off the computational efficiency and the feature representation capability, as well as the object detection accuracy, we configured the dimension of a capsule as 16 for all capsule layers and the regularization factors $\lambda_1=1$, $\lambda_2=1$, and $\lambda_3=1$.

Geospatial object detection

At the test stage, we applied the OA-CapsNet to the test sets of the GOD²18 and the DOTA datasets to evaluate its object detection performance. To provide quantitative evaluations on the object detection results, we adopted the following three commonly used evaluation metrics: precision (P), recall (R), and F_{score} . Precision and recall, respectively, evaluate the performance of an object detection model in distinguishing false alarms and identifying true targets. F_{score} provides an overall performance evaluation by taking into account both the precision and recall measures. These three quantitative evaluation metrics are formally defined as follows:

$$P = \frac{(TP)}{(TP) + (FP)} \quad (9)$$

$$R = \frac{(TP)}{(TP) + (FN)} \quad (10)$$

$$F_{\text{score}} = 2 \times \frac{P \times R}{P + R} \quad (11)$$

where TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively. The geospatial object detection results obtained by the proposed OA-CapsNet on the two test datasets are reported in Table 1 by using the above three quantitative evaluation metrics. In Table 1, ‘‘Average’’ denotes

Table 1. Geospatial object detection results obtained by different methods.

| Method | Dataset | Quantitative evaluation | | |
|------------------|---------------------|-------------------------|--------|-------------|
| | | Precision | Recall | F_{score} |
| OA-CapsNet | GOD ² 18 | 0.9687 | 0.9276 | 0.9477 |
| | DOTA | 0.9563 | 0.9180 | 0.9368 |
| | Average | 0.9625 | 0.9228 | 0.9423 |
| OA-CapsNet-CFA | GOD ² 18 | 0.9522 | 0.9213 | 0.9365 |
| | DOTA | 0.9457 | 0.9121 | 0.9286 |
| | Average | 0.9490 | 0.9167 | 0.9326 |
| OA-CapsNet-SFA | GOD ² 18 | 0.9486 | 0.9192 | 0.9337 |
| | DOTA | 0.9391 | 0.9105 | 0.9246 |
| | Average | 0.9439 | 0.9149 | 0.9292 |
| OA-CapsNet-Light | GOD ² 18 | 0.9259 | 0.8977 | 0.9116 |
| | DOTA | 0.9168 | 0.8913 | 0.9039 |
| | Average | 0.9214 | 0.8945 | 0.9078 |
| MS-VAN | GOD ² 18 | 0.9036 | 0.8769 | 0.8900 |
| | DOTA | 0.8931 | 0.8724 | 0.8826 |
| | Average | 0.8984 | 0.8747 | 0.8863 |
| SOON | GOD ² 18 | 0.9127 | 0.8857 | 0.8990 |
| | DOTA | 0.9032 | 0.8816 | 0.8923 |
| | Average | 0.9080 | 0.8837 | 0.8957 |
| MSCNN | GOD ² 18 | 0.9459 | 0.9173 | 0.9314 |
| | DOTA | 0.9362 | 0.9081 | 0.9219 |
| | Average | 0.9411 | 0.9127 | 0.9267 |
| RFN | GOD ² 18 | 0.9398 | 0.9118 | 0.9256 |
| | DOTA | 0.9315 | 0.9044 | 0.9177 |
| | Average | 0.9357 | 0.9081 | 0.9217 |
| FMSSD | GOD ² 18 | 0.9315 | 0.9052 | 0.9182 |
| | DOTA | 0.9238 | 0.8963 | 0.9098 |
| | Average | 0.9277 | 0.9008 | 0.9140 |
| RECNN | GOD ² 18 | 0.9223 | 0.8946 | 0.9082 |
| | DOTA | 0.9131 | 0.8875 | 0.9001 |
| | Average | 0.9177 | 0.8911 | 0.9042 |

the average performance obtained on the two test datasets. It includes the average precision, average recall, and average F_{score} .

As reported in Table 1, the proposed OA-CapsNet achieved quite promising performances on the two test datasets. An object detection accuracy with a precision, a recall, and an F_{score} of 0.9687, 0.9276, and 0.9477, respectively, was obtained on the GOD²18 dataset. For the DOTA dataset, an accuracy with a precision, a recall, and an F_{score} of 0.9563, 0.9180, and 0.9368, respectively, was achieved in detecting geospatial objects. Comparatively, the DOTA dataset involved more categories of geospatial objects with more varying and complicated conditions, thus, a relatively better performance was obtained on the GOD²18 dataset. Specifically, for each dataset, the precision metric was better than the recall metric. It means that the proposed OA-CapsNet behaved promisingly in distinguishing the true targets and the false alarms, thereby resulting in a low false recognition rate. Overall, the object detection performance was quite competitive in processing the two remarkably challenging datasets. An average object detection accuracy with a precision, a recall, and an F_{score} of 0.9625, 0.9228, and 0.9423, respectively, was achieved on the two test datasets.

The challenging scenarios of these two datasets cover the following aspects: (1) objects with varying scales and orientations, (2) objects with diverse texture properties and spatial distributions, (3) objects with different levels of occlusions, (4) objects contaminated by different levels of shadows, (5) similarities between the objects of interest and the non-targets, (6) illumination condition variations, and (7) complicated environmental conditions. These challenging scenarios impeded the highly accurate recognition of the geospatial objects, and required that the object detection model should be self-adaptive, robust, effective, and transferable enough to correctly identify the existence of objects, accurately locate the positions of objects, effectively distinguish the true targets and false alarms, and applicably handle different image sources. Fortunately, the proposed OA-CapsNet still performed promisingly with a high detection accuracy in processing the geospatial objects of varying conditions in diverse scenarios. The advantageous performance of the proposed OA-CapsNet benefited from the following aspects. First, by employing a capsule feature pyramid network architecture as the backbone, the proposed OA-CapsNet can extract and fuse multiscale and multilevel high-order capsule features to provide a semantically strong feature representation at each scale. Second, by integrating the two types of capsule-based feature attention modules, the proposed OA-CapsNet can highlight channel-wise informative features and focus on class-specific spatial features to further enhance the feature representation quality and robustness. Last but not least, by designing a centre-ness-assisted anchor-free object detection network, the proposed OA-CapsNet can detect arbitrarily-oriented and varying-scale objects.

For visual inspections, Figure 5 also presents a subset of geospatial object detection results from the two test datasets. As observed by the object detection results in Figure 5, the objects of different scales, arbitrary orientations, diverse densities, and varying distributions in different environmental scenarios were effectively recognized. Specifically, for the images containing very high-density and parallel-distributed vehicles and ships, the proposed OA-CapsNet still achieved a competitive detection performance owing to the design of the five-tuple based orientated bounding box representation. In addition, for the images containing different-scale objects, especially the small-size objects (e.g. ships, vehicles, airplanes, storage tanks, etc.), the proposed OA-CapsNet still performed promisingly in correctly identifying and locating these objects due to the semantically strong pyramidal



Figure 5. Illustration of a subset of geospatial object detection results from the GOD²18 and DOTA datasets.



Figure 6. Illustration of some special challenging scenarios of the geospatial objects.

feature representations used for object detection at multiple scales. Moreover, benefited from the high-quality, informative, and object-orientated feature encodings upgraded by the capsule-based channel and spatial feature attention modules at each scale, some geospatial objects partially occluded by the nearby high-rise objects or covered with dark shadows were also effectively detected with a quite low misidentification rate. However, as shown by the green boxes in [Figure 6](#), some geospatial objects were severely occluded by the nearby objects, leading to the extreme incompleteness in the remote sensing images and quite few feature presences in the feature maps. As a result, these objects were failed to be correctly identified. In addition, some land covers exhibited quite similar geometric and textural properties to the geospatial objects. Consequently, they were falsely recognized as the true targets. Moreover, some objects were covered with large-area heavy dark shadows or of

extremely small sizes, making the objects hide into the background. Unfortunately, these objects were treated as the background and failed to be correctly detected. On the whole, the proposed OA-CapsNet achieved an acceptable performance in detecting geospatial objects of different conditions in diverse scenarios.

Ablation studies

As ablation studies, we further examined the performance improvement achieved by integrating the CFA and SFA modules into the pyramidal feature extraction backbone for enhancing the feature representation robustness and informativeness at multiple scales. Specifically, the CFA module functioned to exploit and highlight the channel-wise informative features related to the foreground and weaken the contributions of the background features. The SFA module served to concentrate on the class-specific spatial

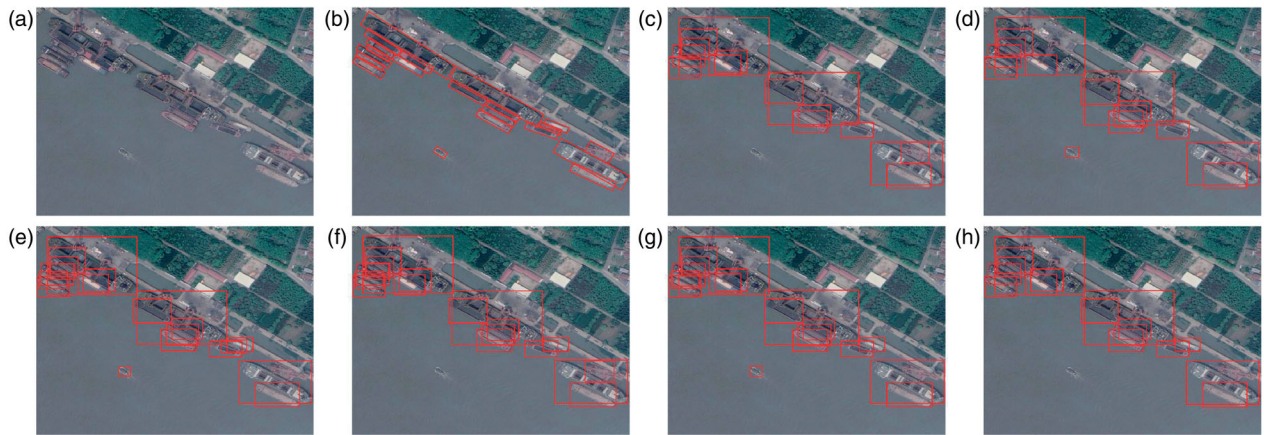


Figure 7. Illustration of ship detection results obtained by different models. (a) Test image, (b) the proposed OA-CapsNet, (c) MS-VAN, (d) SOON, (e) MSCNN, (f) RFN, (g) FMSSD, and (h) RECNN.

features associated with the object regions and suppress the impacts of the background areas. To this end, we constructed three modified networks on the basis of the OA-CapsNet. First, we removed all the SFA modules from the OA-CapsNet, leaving only the CFA modules, and named the resultant network as OA-CapsNet-CFA. Then, we removed all the CFA modules from the OA-CapsNet, leaving only the SFA modules, and named the resultant network as OA-CapsNet-SFA. Finally, we removed all the CFA and SFA modules from the OA-CapsNet, and named the resultant network as OA-CapsNet-Light.

For fair comparisons, the same training sets of the GOD²18 and DOTA datasets and the same training strategy were leveraged to train these three networks. Once the network parameters of these three networks were fine-tuned, we applied them to the test sets of the GOD²18 and DOTA datasets to evaluate their object detection performances. The detailed object detection results measured using the precision, recall, and F_{score} metrics are reported in Table 1. Obviously, without the integration of the CFA and SFA modules, the object detection accuracies of the OA-CapsNet-Light were dramatically degraded on both of the two test datasets. The accuracy degradation with regard to the average F_{score} on the two test datasets was about 0.0345. In contrast, by integrating the CFA or the SFA modules into the multiscale feature extraction backbone, the quality of the output features was significantly upgraded to positively support the objects of interest, therefore, with the high-quality feature representations fed into the object detection heads, the object detection performance was improved by the OA-CapsNet-CFA and OA-CapsNet-SFA. Comparatively, the OA-CapsNet-CFA behaved relatively better than the OA-CapsNet-SFA with a

performance upgradation of about 0.0034 with regard to the average F_{score} . As a conclusion, both of the CFA and SFA modules contributed positively and effectively to the enhancement of the feature representation quality and the improvement of the object detection performance. Therefore, with the integration of the CFA and SFA modules to highlight both the channel-wise informative features and the class-specific spatial features associated with the foreground, the OA-CapsNet showed advantageous performance in processing geospatial objects of varying conditions in diverse scenarios.

Comparative studies

To further evaluate the feasibility and performance of the proposed OA-CapsNet, we conducted a set of comparative studies with the following six recently developed one-stage object detection methods: MS-VAN (Wang et al. 2019), SOON (Qin et al. 2021), MSCNN (Yao et al. 2021), RFN (Zhou, Zhang, Gao, et al. 2020a), FMSSD (Wang et al. 2020), and RECNN (Lei et al. 2020). Specifically, the MS-VAN leveraged multiscale features and attention mechanisms to produce high-quality feature encodings used for object detection. The SOON exploited spatial properties via a receptive field enhancement module to protrude the feature semantics of small-size objects. The MSCNN designed an effective feature extraction backbone to obtain strong feature representations at multiple scales. The RFN focused on the extraction of orientation-aware feature maps to boost the recognition of arbitrarily-orientated objects. The FMSSD extracted and fused spatial contextual features in both multiple scales and the same scales to well handle varying-size geospatial objects. The RECNN introduced a saliency

constraint and multilayer fusion strategy to strengthen the feature saliencies of the object regions.

In our experiments, for fair comparisons, the same training sets of the GOD²18 and the DOTA datasets were used to train these models. Once the models were constructed, we applied them to the test sets of these two datasets to evaluate their object detection performances. The quantitative evaluations obtained by these models on the object detection results are reported in Table 1. As reflected in Table 1, the MSCNN, RFN, and FMSSD showed superior performances than the RECNN, MS-VAN and SOON. Specifically, the object detection performance of the MSCNN was higher than that of the MS-VAN by about 0.0404 with regard to the average F_{score} . The advantageous performances of the MSCNN, RFN, and FMSSD benefited from the exploitation of effective mechanisms to take advantage of multiscale or multi-orientated features to improve the feature representation informativeness and robustness. Thus, they performed promisingly in correctly recognizing the geospatial objects of varying conditions in diverse scenarios. Comparatively, by designing a capsule feature pyramid network architecture as the feature extraction backbone and integrating the two types of capsule feature attention modules to provide multi-scale semantically strong and informative features, as well as the effective centreness-assisted anchor-free object detection strategy, the proposed OA-CapsNet showed competitive and superior performance over the six compared methods. For visual inspections and comparisons, Figure 7 presents some examples of ship detection results obtained by using these models. As shown by Figures 7c–7h, some ships of extremely small sizes and some ships distributed parallelly and closely were not correctly recognized. In contrast, all the ships of varying conditions were successfully detected by the proposed OA-CapsNet. Thus, through contrastive analysis, we concluded that the proposed OA-CapsNet provided a feasible and effective solution to geospatial object detection tasks.

Conclusion

This paper has presented a novel one-stage anchor-free capsule network, named OA-CapsNet, for geospatial object detection from remote sensing images. Formulated with a capsule feature pyramid network architecture as the backbone, the proposed OA-CapsNet can extract and fuse multilevel and multiscale high-order capsule features to provide a high-quality, semantically strong feature encoding at

each scale. Integrated with two types of capsule feature attention modules, the proposed OA-CapsNet performed effectively in highlighting the channel-wise informative features and focusing on the class-specific spatial features to further enhance the feature representation quality and robustness. Designed with a centreness-assisted anchor-free object detection strategy, the proposed OA-CapsNet served to effectively recognize arbitrarily-orientated and diverse-scale geospatial objects. The proposed OA-CapsNet has been intensively evaluated on two large remote sensing image datasets toward geospatial object detection. Quantitative evaluations showed that a competitive overall performance with a precision, a recall, and an F_{score} of 0.9625, 0.9228, and 0.9423, respectively, was achieved in handling geospatial objects of varying conditions in diverse environmental scenarios. Comparative studies with a set of recently developed deep learning methods also confirmed the applicability and effectiveness of the proposed OA-CapsNet in geospatial object detection tasks. However, due to severe occlusions, poor illumination conditions, and extremely small sizes of some geospatial objects, it is still challengeable to fulfill high-performance geospatial object detection. In our future work, we will develop part-based models to improve the detection of occluded objects, design more powerful feature attention mechanisms to highlight the low-contrast objects, and exploit high-resolution network architectures or super-resolution techniques to effectively handle small-size objects.


Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the [National Natural Science Foundation of China] under Grants [62076107], [51975239], [41971414], and [41671454]; by the [Six Talent Peaks Project in Jiangsu Province] under Grant [XYDXX-098]; and by the [Natural Science Foundation of Jiangsu Province] under Grant [BK20191214].

ORCID

Yongtao Yu  <http://orcid.org/0000-0001-7204-9346>
 Haiyan Guan  <http://orcid.org/0000-0003-3691-8721>
 Dilong Li  <http://orcid.org/0000-0002-5826-5568>
 Jonathan Li  <http://orcid.org/0000-0001-7899-0049>

References

- Bao, S., Zhong, X., Zhu, R., Zhang, X., Li, Z., and Li, M. 2019. "Single shot anchor refinement network for oriented object detection in optical remote sensing imagery." *IEEE Access.*, Vol. 7: pp. 87150–87161. doi:10.1109/ACCESS.2019.2924643.
- Chen, S., Zhan, R., and Zhang, J. 2018. "Geospatial object detection in remote sensing imagery based on multiscale single-shot detector with activated semantics." *Remote Sensing*, Vol. 10(No. 6): pp. 820. doi:10.3390/rs10060820.
- Courtrai, L., Pham, M.T., and Lefèvre, S. 2020. "Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks." *Remote Sensing*, Vol. 12(No. 19): pp. 3152. doi:10.3390/rs12193152.
- Cozzolino, D., Martino, G.D., Poggi, G., and Verdoliva, L. 2017. "A fully convolutional neural network for low-complexity single-stage ship detection in Sentinel-1 SAR images." Paper presented at the IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, USA, July 2017.
- Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. 2019. "Dual attention network for scene segmentation." Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, June 2019.
- Gong, Y., Xiao, Z., Tan, X., Sui, H., Xu, C., Duan, H., and Li, D. 2020. "Context-aware convolutional neural network for object detection in VHR remote sensing imagery." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 58(No. 1): pp. 34–44. doi:10.1109/TGRS.2019.2930246.
- Hu, Y., Li, X., Zhou, N., Yang, L., Peng, L., and Xiao, S. 2019. "A sample update-based convolutional neural network framework for object detection in large-area remote sensing images." *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(No. 6): pp. 947–951. doi:10.1109/LGRS.2018.2889247.
- Hu, J., Shen, L., and Sun, G. 2018. "Squeeze-and-excitation networks." Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, June 2018.
- Lei, J., Luo, X., Fang, L., Wang, M., and Gu, Y. 2020. "Region-enhanced convolutional neural network for object detection in remote sensing images." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 58(No. 8): pp. 5693–5702. doi:10.1109/TGRS.2020.2968802.
- Li, Q., Mou, L., Xu, Q., Zhang, Y., and Zhu, X.X. 2019. "R³-Net: A deep network for multioriented vehicle detection in aerial images and videos." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57(No. 7): pp. 5028–5042. doi:10.1109/TGRS.2019.2895362.
- Li, Y., Pei, X., Huang, Q., Jiao, L., Shang, R., and Marturi, N. 2020. "Anchor-free single stage detector in remote sensing images based on multiscale dense path aggregation feature pyramid network." *IEEE Access.*, Vol. 8: pp. 63121–63133. doi:10.1109/ACCESS.2020.2984310.
- Li, K., Wan, G., Cheng, G., Meng, L., and Han, J. 2020. "Object detection in optical remote sensing images: A survey and a new benchmark." *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 159: pp. 296–307. doi:10.1016/j.isprsjprs.2019.11.023.
- Lin, T.Y., Goyal, P., Girshick, R., He, K., and Dollár, P. 2017. "Focal loss for dense object detection." Paper presented at the IEEE International Conference on Computer Vision, Venice, Italy, October 2017.
- Liu, W., Ma, L., Wang, J., and Chen, H. 2019. "Detection of multiclass objects in optical remote sensing images." *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(No. 5): pp. 791–795. doi:10.1109/LGRS.2018.2882778.
- Liu, Q., Xiang, X., Yang, Z., Hu, Y., and Hong, Y. 2021. "Arbitrary direction ship detection in remote-sensing images based on multitask learning and multiregion feature fusion." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59(No. 2): pp. 1553–1564. doi:10.1109/TGRS.2020.3002850.
- Mandal, M., Shah, M., Meena, P., Devi, S., and Vipparthi, S.K. 2020. "AVDNet: A small-sized vehicle detection network for aerial visual data." *IEEE Geoscience and Remote Sensing Letters*, Vol. 17(No. 3): pp. 494–498. doi:10.1109/LGRS.2019.2923564.
- Mekhalfi, M.L., Bejiga, M.B., Soresina, D., Melgani, F., and Demir, B. 2019. "Capsule networks for object detection in UAV imagery." *Remote Sensing*, Vol. 11(No. 14): pp. 1694. doi:10.3390/rs11141694.
- Pham, M.T., Courtrai, L., Friguet, C., Lefèvre, S., and Baussard, A. 2020. "YOLO-fine: One-stage detector of small objects under various backgrounds in remote sensing images." *Remote Sensing*, Vol. 12(No. 15): pp. 2501. doi:10.3390/rs12152501.
- Qin, H., Li, Y., Lei, J., Xie, W., and Wang, Z. 2021. "A specially optimized one-stage network for object detection in remote sensing images." *IEEE Geoscience and Remote Sensing Letters*, Vol. 18 (No. 3): pp. 401–405. doi:10.1109/LGRS.2020.2975086.
- Rabbi, J., Ray, N., Schubert, M., Chowdhury, S., and Chao, D. 2020. "Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network." *Remote Sensing*, Vol. 12(No. 9): pp. 1432. doi:10.3390/rs12091432.
- Rajasegaran, J., Jayasundara, V., Jayasekara, S., Jayasekara, H., and Seneviratne, S.R.R. 2019. "DeepCaps: Going deeper with capsule networks." Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA.
- Rezatofighi, H., Tsoi, N., Gwak, J.Y., Sadeghian, A., Reid, I., and Savarese, S. 2019. "Generalized intersection over union: A metric and a loss for bounding box regression." Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, June 2019.
- Sabour, S., Frosst, N., and Hinton, G.E. 2017. "Dynamic routing between capsules." Paper presented at the 31st Conference on Neural Information Processing Systems, Long Beach, USA.
- Shi, F., Zhang, T., and Zhang, T. 2020. "Orientation-aware vehicle detection in aerial images via an anchor-free object detection approach." *IEEE Transactions on Geoscience and Remote Sensing*. Advance online publication. doi:10.1109/TGRS.2020.3011418.
- Tang, G., Liu, S., Fujino, I., Claramunt, C., Wang, Y., and Men, S. 2020. "H-YOLO: A single-shot ship detection

- approach based on region of interest preselected network.” *Remote Sensing*, Vol. 12(No. 24): pp. 4192. doi:[10.3390/rs12244192](https://doi.org/10.3390/rs12244192).
- Tang, T., Zhou, S., Deng, Z., Lei, L., and Zou, H. 2017. “Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks.” *Remote Sensing*, Vol. 9(No. 11): pp. 1170. doi:[10.3390/rs9111170](https://doi.org/10.3390/rs9111170).
- Tian, S., Kang, L., Xing, X., Li, Z., Zhao, L., Fan, C., and Zhang, Y. 2020. “Siamese graph embedding network for object detection in remote sensing images.” *IEEE Geoscience and Remote Sensing Letters*. Advance online publication. doi:[10.1109/LGRS.2020.2981420](https://doi.org/10.1109/LGRS.2020.2981420).
- Wang, C., Bai, X., Wang, S., Zhou, J., and Ren, P. 2019. “Multiscale visual attention networks for object detection in VHR remote sensing images.” *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(No. 2): pp. 310–314. doi:[10.1109/LGRS.2018.2872355](https://doi.org/10.1109/LGRS.2018.2872355).
- Wang, P., Sun, X., Diao, W., and Fu, K. 2020. “FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery.” *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 58(No. 5): pp. 3377–3390. doi:[10.1109/TGRS.2019.2954328](https://doi.org/10.1109/TGRS.2019.2954328).
- Wu, X., Hong, D., Ghamisi, P., Li, W., and Tao, R. 2018. “MsRi-CCF: Multi-scale and rotation-insensitive convolutional channel features for geospatial object detection.” *Remote Sensing*, Vol. 10(No. 12): pp. 1990. doi:[10.3390/rs10121990](https://doi.org/10.3390/rs10121990).
- Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. 2018. “DOTA: A large-scale dataset for object detection in aerial images.” Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, June 2018.
- Yao, Q., Hu, X., and Lei, H. 2021. “Multiscale convolutional neural networks for geospatial object detection in VHR satellite images.” *IEEE Geoscience and Remote Sensing Letters*, Vol. 18 (No. 1): pp. 23–27. doi:[10.1109/LGRS.2020.2967819](https://doi.org/10.1109/LGRS.2020.2967819).
- Yu, Y., Guan, H., Li, D., Gu, T., Tang, E., and Li, A. 2020. “Orientation guided anchoring for geospatial object detection from remote sensing imagery.” *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 160: pp. 67–82. doi:[10.1016/j.isprsjprs.2019.12.001](https://doi.org/10.1016/j.isprsjprs.2019.12.001).
- Zhang, Z., Liu, Y., Liu, T., Lin, Z., and Wang, S. 2020. “DAGN: A real-time UAV remote sensing image vehicle detection framework.” *IEEE Geoscience and Remote Sensing Letters*, Vol. 17(No. 11): pp. 1884–1888. doi:[10.1109/LGRS.2019.2956513](https://doi.org/10.1109/LGRS.2019.2956513).
- Zhang, X., Wang, G., Zhu, P., Zhang, T., Li, C., and Jiao, L. 2020. “GRS-Det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images.” *IEEE Transactions on Geoscience and Remote Sensing*. Advance online publication. doi:[10.1109/TGRS.2020.3018106](https://doi.org/10.1109/TGRS.2020.3018106).
- Zheng, Z., Zhong, Y., Ma, A., Han, X., Zhao, J., Liu, Y., and Zhang, L. 2020. “HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery.” *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 166: pp. 1–14. doi:[10.1016/j.isprsjprs.2020.04.019](https://doi.org/10.1016/j.isprsjprs.2020.04.019).
- Zhou, K., Zhang, Z., Gao, C., and Liu, J. 2020a. “Rotated feature network for multiorientation object detection of remote-sensing images.” *IEEE Geoscience and Remote Sensing Letters*, Vol. 18(No. 1): pp. 33–37. doi:[10.1109/LGRS.2020.2965629](https://doi.org/10.1109/LGRS.2020.2965629).
- Zhou, C., Zhang, J., Liu, J., Zhang, C., Shi, G., and Hu, J. 2020b. “Bayesian transfer learning for object detection in optical remote sensing images.” *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 58(No. 11): pp. 7705–7719. doi:[10.1109/TGRS.2020.2983201](https://doi.org/10.1109/TGRS.2020.2983201).
- Zhuang, S., Wang, P., Jiang, B., Wang, G., and Wang, C. 2019. “A single shot framework with multi-scale feature fusion for geospatial object detection.” *Remote Sensing*, Vol. 11(No. 5): pp. 594. doi:[10.3390/rs11050594](https://doi.org/10.3390/rs11050594).