Capsule-inferenced Object Detection for Remote Sensing Images

Yingchao Han, Weixiao Meng, *Senior Member, IEEE*, and Wei Tang, *Member, IEEE*

***Abstract*—Frequent and accurate object detection based on remote sensing images can effectively monitor dynamic objects on the earth's surface. While the detection transformer (DETR) offers a simple encoder–decoder structure and a direct set prediction approach to object detection, it falls short in complex remote sensing scenes where entity information and relative positions between objects are critical to target reasoning. Notably, the DETR model's feedforward neural network (FFN) relies on weighted summation for target reasoning, disregarding interactive feature information, which is a major factor affecting detection effectiveness. To address these shortcomings, in this article, we propose a DETR-based detection model called (CI_DETR), which uses capsule inference to improve remote sensing object detection. Our approach adds a multilevel feature fusion module to the DETR network, allowing the network to learn how to spatially alter features at different levels, preserving only beneficial information for combination. In addition, we introduce a capsule reasoning module to mine entity information during inference and more effectively model the hierarchical correlation of internal knowledge representation in the neural network, consistent with the thinking model of the human brain. Lastly, we employ a sausage model to measure the similarities and differences of capsules, projecting them onto a curved surface for nonlinear function approximation and maximum preservation of the local responsiveness of capsule entities. Our experiments on public datasets confirm the superior detection performance of our proposed algorithm relative to many current detectors.**

***Index Terms*—CapsNet, object detection, remote sensing image, transformer.**

I. Introduction

OBJECT detection plays a critical role in remote sensing processing technology. Its applications [1], [2], [3], [4], [5] have been employed in various fields, including military investigations, environmental monitoring, dangerous target tracking, urban traffic management, and geographic information services.

A wide variety of object recognition [6], [7], tracking [8], [9], [10], and detection [3], [4], [11] algorithms have been put forward by the computer vision community in the past decades, and their performance has been rather spectacular. In deep learning-based object detection, selecting the appropriate candidate bounding box/anchor from the global scene is crucial. Furthermore, the characteristics of the respective candidate must be extracted using convolutional neural network [6], [12], [13], [14]. Next, they are fed into a classifier to determine whether the bounding box contains an object and to detect the type of object [15]. On that basis, the position and categorization of objects are determined. The candidate bounding box/anchor of the object needs to be chosen from the global scene, followed by the extraction of the convolutional neural network [6], [12], [13], [14] characteristics of the respective candidate. Subsequently, they are fed to a classier to determine whether the bounding box includes the object as well as detect the object types [15]. Accordingly, the position and categorization of objects are determined.

Traditional natural images are typically taken from a low-angle perspective and depict a clear scene with a prominent subject. However, earth observation images are captured using satellite or aerial photography from a top-down perspective. This creates challenges due to the varying number, scale, orientation, and geometric distortion of geospatial objects, resulting in an exponentially increased number of potential bounding boxes to be searched [16], [17]. Furthermore, geographical objects are typically dispersed in a heterogeneous manner and combined with cluttered backgrounds. As a result, the conventional convolutional method with a narrow receptive field cannot comprehend the global context of a geographic image. [18], [19]. As a result, achieving frequent and reliable recognition of geographical items from earth observation remains a significant challenge.

Over the past few years, significant advancements have been made in transformer-based object detection methods [20], [21]. By adopting self-attention layers for long-distance dependency modeling, this method replaces the convolutional layers locality modeling and represents the global interaction between heterogeneously distributed objects [22]. It also exhibits significant abilities to distinguish their types and locations from cluttered backgrounds [23], [24]. The transformer-based method is capable of reformulating the detection problem as a paradigm in terms of disorder set prediction and matching, which can thus be capable of automatically matching the prediction (type and bounding box) with its ground truth as well as reducing the demands of hand-designed region-proposal/ anchors [20], inconsistent with the conventional object detection manner employing pairwise prediction and ground truth denied by the region-proposal/anchors for training. Accordingly, this method

Manuscript received 20 February 2023; revised 15 March 2023; accepted 30 March 2023. Date of publication 13 April 2023; date of current version 15 June 2023. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 91438203. *(Corresponding author: Weixiao Meng.)*

Yingchao Han and Weixiao Meng are with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: 21b905067@stu.hit.edu.cn; wxmeng@hit.edu.cn).

Wei Tang is with the State Key Laboratory of Space-Earth Integration Information Technology, Beijing Fifth Academy of Aerospace, Beijing 100048, China (e-mail: tgwi@foxmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3266794

has the potential to be employed for geospatial object detection in earth observation images.

Matching entity information to features and relative position information between entities is critical for target reasoning in the detection of remote sensing targets, particularly for detecting multi-scale and dense objects. However, in the transformer-based object detection method (such as DETR), the feedforward neural network (FFN) only achieves target information inference through weighted summation and does not consider the interactive information between features, which becomes a major factor affecting the detection effect. In contrast, multi-scale and dense objects are easier to detect by humans as neuroscience suggests that anything humans see converges in various ways into a continuous attractor, which may manifest as a curve, a surface, or a hypersurface [25]. In accordance with the above assumption, coverage learning can be performed on samples through continuous topological geometry in high-dimensional spaces [26]. In addition, starting from the thinking model of the human brain, Hinton et al. [27] built capsules to describe entities and their attributes and better model the hierarchical relationship and the transfer method of internal knowledge representation in the network using the dynamic routing information transfer method. For object detection tasks, the design of such a hypersurface will significantly affect the detection effect.

The current capsule networks face limitations that restrict their extensive use on complex datasets. Firstly, dynamic routing has been found to be computationally expensive, and having many layers can result in a significant increase in training and inference time, making it infeasible for large and complex datasets. Secondly, the focus of the squash activation function, and its variants, is mainly on preserving the vector orientation, while capsule activation is primarily aimed at creating a function of capsule-scale activation.

Drawing inspiration from human brain neuroscience, we propose a novel Transformers object-detection model that combines capsule inference and multi-scale feature enhancement (CI_DETR). CI_DETR comprehensively adopts the multi-scale features generated by the backbone network, which contain both semantic and spatial information, to detect targets of varying scales. To overcome the limitation of FFN in not accounting for mutual information between features and to approximate the thinking model of the human brain, we simulate the layering and transfer of internal knowledge representation in the network to build a capsule inference module. We combine capsule construction and information routing to reason about target types and locations.To enhance the nonlinear expressiveness of capsule inference, we introduce non-linear sausage metrics to replace squash activation. This new approach approximates non-linear mapping with arbitrary precision and enables the capsules to represent entities with low correlation. In summary, our proposed CI_DETR model addresses the limitations of dynamic routing and squash activation and enhances the multi-scale feature detection of complex datasets through capsule.

The novelties of our work are presented in threefolds.

1) The present study focuses on developing a brain-inspired object detection framework for remote sensing images, thereby effectively detecting objects with the consideration of the targets unique characteristics in images.
2) To more accurately simulate the layering and transmission of internal knowledge representation in the human brain, we have developed a capsule reasoning module that integrates capsule construction and information routing. This module enables the reasoning about target categories and locations by incorporating relative position information and category associations of the target during the reasoning process. By doing so, the capsule reasoning module is capable of capturing global information and enhancing object detection performance. In essence, our capsule reasoning module is intended to closely emulate the way the human brain processes information and represents knowledge, ultimately leading to improved object detection performance.
3) Numerous experiments are performed to confirm the effectiveness of the proposed method in relative to the current research.

The rest of this article is organized as follows. In Section II, we briefly show the works related to object detection, DETR with an end-to-end objective, and capsule network. In Section III, we display the model architecture of the proposed capsule-inference object detection algorithm. In Section IV, we focus on discussing the performance of the proposed method and also compare it with that of state-of-the-art object detection methods on DIOR and HRRSD datasets. Finally, Section V concludes this article.

II. Related Work

As displayed in the present section, we show a brief introduction to object detection networks, DETR with an end-to-end objective, and capsule networks. The above research has significantly contributed to the proposed method.

A. Object Detection

The object detection task refers to the determination of an object and the identification of its location in the image or video. Benefiting from deep neural networks, numerous object detection methods have obtained obvious progress over the past few years. Faster R-CNN [28] represents an end-to-end detection method which can replace selective searching with a novel region proposal network. SSD [29] can predict several bounding boxes at various scales and aspect ratios from several feature layers. RetinaNet [30] employs a focus loss function to solve the type imbalance problem of single-stage detectors. FCOS [31] refers to an anchorless object detector introducing centrality to further enhance detection performance. Other object detection algorithms [32], [33], [34] employ one or several points to represent an object, thus ensuring the balance between speed and accuracy. Detection TRansformer(DETR) [20] is a recently suggested end-to-end object detection system that adopts Hungarian matching for label assignment. Although it achieves equivalent performance to Faster R-CNN, its detection performance on small objects is inferior, and it has a low convergence rate.

B. End-to-End DETR

The DEtection TRansformer (DETR) [20] introduces the first technique with an end-to-end optimization target for set prediction, in contrast to the aforementioned widely used object detectors. It specifically uses a bipartite matching method to formulate the loss function. Encoder–decoder transformer [35] is the structure used by DETR [20], which is based on CNN. The paper proposes processing flattened deep features from the CNN backbone with a Transformer encoder component. The decoder part, which is non-autoregressive, uses the encoder outputs and previously learned object query vectors to predict category labels and bounding boxes as the detection result. The cross-attention module in the decoder plays a crucial role in attending to various areas in the image for different object queries. For those not familiar with transformers, we recommend referring to the appendix. Thanks to the ability of the self-attention component to learn the removal of repeated detections, specifically with the Hungarian loss incentivizing one target per object in DETR, the attention mechanism eliminates the need for NMS post-processing.

In parallel with our research, various DETR versions have been put out to enhance the effectiveness and precision of its training. According to Deformable DETR [21], the idea of deformable convolution and attention modules should be combined in order to construct a sparse attention mechanism on multi-level feature maps. When DETR is tailored for downstream tasks, UP-DETR [36] uses an unsupervised pre-training job called random query patch detection in order to boost its performance. Compared to these work, we improve the defect of missing relative position information in the process of DETR detection head presence detection reasoning, and enhance the performance of DETR in remote sensing target detection.

C. Object Detection in Remote Sensing Images

Based on the fast advancement of the aforementioned algorithms, other approaches for object recognition in remote-sensing images have been presented ([37], [38], [39], [40], [41], [42], [43], [44], [45], [46]). In order to accomplish rotation-invariant detection, several research projects concentrate on identifying arbitrarily oriented objects using horizontal bounding box annotations. For instance, Cheng et al. ([37]) suggested a CNN that is rotation-invariant (RICNN), including a new rotation-invariant layer. To address the issue of substantial disparities in target scales, a great deal of research focuses on optimizing feature pyramid networks in order to extract effective multiscale features. The ABNet ([38]) detection approach, for instance, creates an adaptive feature pyramid network to adaptively fuse multilevel scale features in the feature pyramid network using channel attention and spatial attention processes. To solve the difficulty that comparable objects have different forms and it is difficult to effectively extract common characteristics for unified representation, several research works have improved anchor frame parameter optimization. For instance, FFA ([39]) and DRBox([40]) employ anchor boxes of various sizes, aspect ratios, and angles to increase the algorithm's capacity to generalize to object shapes.

D. Capsule Network

The CapsNet Network [27], a representative of bionics, proposed by Hinton is specially designed for CNN-based feature extraction and has aroused wide attention from artificial intelligence researchers. The transmission and operation logic of the capsule network is more consistent with the way neurons in the human brain work. Capsules have a diverse range of capabilities and can exhibit various properties. For example, different areas of the human brain are responsible for different tasks. As our understanding of the human brain deepens and continues to accumulate through neuroscience study, we can understand capsules as groups of neurons whose activity vectors represent the instantiated parameters of an entity with a specific type.

The CapsNet Network (CapsuleNet) is an innovative approach for implementing the capsule concept. Through the introduction of the dynamic routing algorithm and the squash activation function, CapsuleNet applies vector-output capsules as its fundamental unit rather than scalar-output features

$$squash\left(s_j\right) = \frac{\left\|s_j\right\|^2}{1+\left\|s_j\right\|^2}\frac{s_j}{\left\|s_j\right\|} \tag{1}$$

where s_j denotes an activation capsule. Nevertheless, CapsuleNet shows a room for development. Besides, the number of parameters of CapsuleNet is shown to be much larger when compared with that of comparable performance CNN-based models. In addition, the dynamic routing is an iterative process.

The result of the above analysis suggests that the CapsuleNet has high performance in the mining of entity information, the interaction between entities, as well as the reasoning of the target. In the task of object detection, how to detect implicitly defined entities in a limited conditional domain and draw the entity's feature information (e.g., the entity's location, type, pose information, etc.) in real-time is of critical significance. Thus, capsule networks are expected to be vital in object detection tasks.

III. Methodology

Fig. 1 displays the overall framework of the CI_DETR method, following the main encoder–decoder architecture of DETR. Unlike DETR, CI_DETR first adopts a backbone network with feature pyramid network (FPN [47]) for extracting multiscale features from images. Next, a multilevel feature fusion method is adopted for enhancing the small-scale feature information to address the issue of the poor detection performance of DETR for small objects. After enhancing the features, we pass them through the Transformers encoder and decoder to obtain the feature representation for each object. To address the limitation of DETR's FFN, which does not consider the interaction between features, we replace it with a capsule reasoning module based on super sausage metrics. This module uses dynamic routing information transmission and feature correlations to achieve object representation and label prediction. Additionally, we introduce a hyper-sausage measurement model with strong non-linear ability to create a more flexible hypersurface that improves the model's expression ability for objects.

A. Multilevel Feature Fusion

DETR's large input feature maps can increase the complexity of model training, while small input feature maps may omit target information. Thus, we propose a multi-level feature fusion

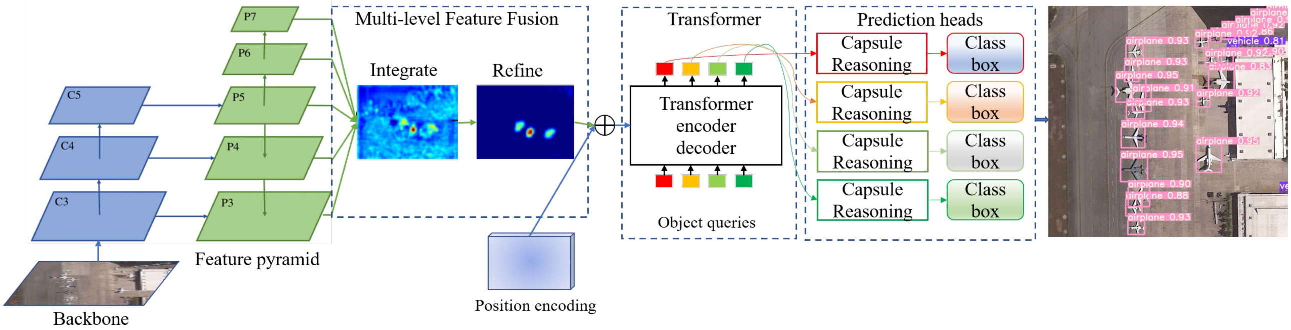


Fig. 1. Workflow of the proposed CI_DETR object detection network.

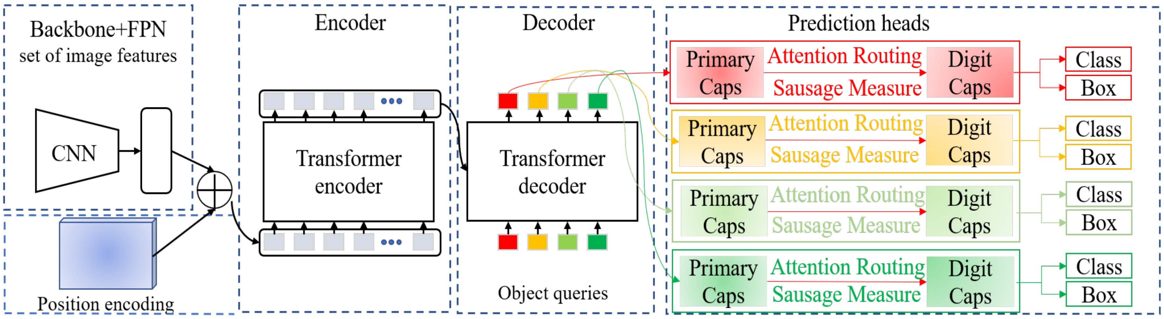


Fig. 2. Realization principle of target detection based on capsule reasoning module in CI_DETR model.

method that integrates and refines features at all FPN levels to balance semantic features. This approach utilizes the benefits of multi-scale feature information and avoids over-burdening model training. The mask-guided module structure, modeled after the Balanced FPN concept in Libra RCNN [48], comprehensively utilizes information from all FPN scale features and achieves balanced feature learning through scale rescaling and feature enrichment.

Step 1. Rescaling: To facilitate subsequent feature aggregation, P_5 feature is selected as the standard, and the multiscale features of FPN are scaled to the same size through upsampling and max pooling operations. To fuse the effective information on the feature layers of different levels, the averaging operation is adopted to perform feature fusion. The fused normalized feature F^{scale} is expressed by the following:

$$F^{Scale} = \frac{1}{5}\sum_{i=3}^{7}\hat{P}_i \tag{2}$$

where $\hat{P}_i$ is the feature generated by feature P_i after feature rescaling.

Step 2. Enriching: To make the integrated features more discriminative, Gaussian nonlocal attention [49] is adopted to refine the feature F^{scale}. Feature refinement can be denoted as

$$U = NL\left(F^{Scale}\right) = NL\left(\frac{1}{5}\sum_{i=3}^{7}\hat{P}_i\right) \tag{3}$$

where U represents the refined features, and $NL(\cdot)$ represents the Gaussian nonlocal attention module. The Gaussian nonlocal attention module can introduce global context information to enhance the expressive power of feature U.

B. Capsule Reasoning

In the DETR model, the FFN only reasons target information through weighted summation, without considering the interaction between features, which has a significant impact on detection effectiveness. To address this issue, we replace the FFN with a capsule inference module during inference on the DETR results. Figure 2 presents the schematic diagram of the capsule inference module for object detection. To construct the basic capsule for Transformer output, we capture the capsule and its corresponding feature dimension information. Instead of a computationally expensive dynamic routing, we adopt self-attention routing to provide an information feed-forward mechanism, which presents object entity attributes more accurately and completely. The Capsule reasoning detects objects (categories and bounding boxes) or 'no object' classes. Finally, the CI_DETR trains the model by utilizing the Hungarian matching algorithm between the labeled and predicted object box. The specific implementation process is illustrated below:

Step 1: We adopt the ResNet backbone network to extract image features and positional encoding, generating a batch of serialized data for CI_DETR. In the encoder stage, attention mechanism extracts features from the serialized data. In the decoder stage, N random initialization vectors are input, and each object query focuses on a different position of the image. After decoding, N vectors are generated, each corresponding to a detected target.

Algorithm 1: Attention Routing using Scalar Product.

```
Input parameters capsule $P_n^l$ from layer $l$;
Output the Digicaps: $V_j$;
Affining transformation for all $P_n^l$:
    $\hat{P}^l_{(n^l,n^{l+1},d^{l+1})} = P_n^{Tl} \times W^l_{(n^l,n^{l+1},d^l,d^{l+1})}$
Calculating self-attention weights:
    $A^l_{(n^l,n^l,n^{l+1})} = \frac{\hat{P}^l_{(n^l,n^{l+1},d^{l+1})\times} \hat{P}^{Tl}_{(n^l,n^{l+1},d^{l+1})}}{\sqrt{d^l}}$
Adopting softmax for Calculating Weights $C$:
    $C^l_{(n^l,n^{l+1},d^{l+1})} = \frac{\exp\left(\sum_{n^l} A^l_{(n^l,n^l,n^{L+1})}\right)}{\sum_{n^{l+1}} \exp\left(\sum_{n^l} A^l_{(n^l,n^l,n^{l+1})}\right)}$
For all capsule $j$ in $l+1$:
    $s_n^{l+1} = \hat{P}^l_{(n^l,n^{l+1},d^{l+1})} \times \left(C^l_{(n^l,n^{l+1})} + B^l_{(n^l,n^{l+1})}\right)$
Compressing the capsule length to between 0 and 1:
    $V_j\,(x_{1\ldots}x_m) \leftarrow sausage\,(s_n^{l+1})$
return $V_j$;
```

Step 2: The feature representation obtained above is converted into a capsule to obtain the initial capsule $P^l_{w_i,h_i,n_id_i}$, in which w_i, h_i, d_i, and n_i suggest the spatial width axis, spatial height axis, capsule dimension axis, and capsule atoms axis, respectively. d_i is adopted to represent the pose, texture, orientation, etc., of the capsule.

Step 3: Self-attention routing. Dynamic routing generated digital capsules lack long-range dependency information between target features, which reduces system robustness. To address this, we replace dynamic routing and squash activation functions with self-attention routing and sausage activation. The batch dot product method transforms dynamic routing to self-attention routing. This enables digital capsules to gather local surrounding information on object features and long-range model dependencies to obtain global information. See Algorithm 1 for specific calculation details.

Algorithm 1 illuminates how the data stream flows in a digital capsule. where $\hat{P}^l_{(n^l,n^{l+1},d^{l+1})}$ includes all predictions of lth capsules. In fact, each n^l capsule, through of the weight matrix, can predict the properties of all n^{l+1} capsules. The term $\sqrt{d^l}$ stabilizes training and contributes to keeping a balance between coupling coefficients and log priors. $B^l_{(n^l,n^{l+1})}$ represents the log priors matrix including all weights discriminatively learned at the same time as all the other weights. In addition, $C^l_{(n^l,n^{l+1})}$ refers to the matrix including all coupling coefficients yielded by the self-attention algorithm. $sausage(\cdot)$ indicates sausage metrics, which is adopted for replacing the squash activation function to calculate the probability value of each attribute of each target capsule in the target as the foundation for judging the target category.

Step 4: Get digital capsule V_j, the Capsule reasoning can predict the normalized center coordinates, height as well as the width of the box w.r.t. the input image, and the linear layer can predict the class label based on a Softmax function. Because we are capable of predicting a fixed-size set of bounding boxes, which is often much larger than the actual number of objects of interest in an image, an additional particular class label ⊘ can be employed to indicate that no object can be identified within a slot. In addition, the class makes a similar role to the background class in the standard object detection approaches.

Step 5: Auxiliary decoding losses: We apply auxiliary losses [50] during training to help output the correct number of objects of each class. We add prediction Capsule reasoning and Hungarian loss after each decoder layer. All prediction Capsule reasonings share the same parameters.

C. Nonlinear Sausage Metrics

Dynamic routing represents an unsupervised algorithm to find a centroidlike output capsule of the prediction capsules. As a result, the squash activation function and its variant 2 [51] concentrate on preserving a capsules orientation

$$squash\ variant\,(s_j) = \left(1 - \frac{1}{\exp\left(\|s_j\|\right)}\right)\frac{s_j}{\|s_j\|}. \quad (4)$$

In this study, we focus on capsule-wise operations without preserving orientation. Capsule activation performs an affine transformation on capsules and applies an element-wise activation function. Capsules on the same channel share parameters for the affine transformation, mapping them to the same feature space. This operation is parameter-efficient, but it cannot preserve vector orientation. However, it is compatible with parameterizing the routing process through attention routing, as capsule activation applies a non-linear transformation to a linear combination of the prediction capsules.

To address these issues, we were inspired by the non-linear sausage measure [52]. We replace the CapsNet squash activation function with the sausage measure, which can approximate non-linear mapping with arbitrary precision and has locally responsive properties.

We determine the similarity between information within a capsule and the difference between information across capsules using the distribution probabilities of upper-level capsules routed from lower-level ones. A topological product on a hypersphere of radius "r" expands the vector indicated by the capsule into a sausage area that employs nonlinear activations to determine the capsule's local responsiveness. Sausage refers to a one-dimensional manifold containing geometry based on covering learning. Fig. 3 illustrates the sausage measurement concept's schematic design. We can mine the submanifold distribution by using convolutional networks to extract low-level features and combining them with the manifold distribution of classes to produce raw data. The sausage's parameters are then

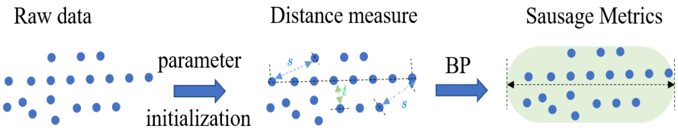


Fig. 3. Schematic diagram of sausage metric model [52].

initialized to determine the Euclidean distance. By employing the non-linear sausage measure, we enhance the capsules' representation and can accommodate low-correlation entities.

Covering learning of complicated dispersed data may be accomplished using the superposition of sausage units. The following depicts the specific implementation.

Step 1: Determine the vector differences between the expected vector and the manifold yielded by the two sausage cores as follows:

$$s = \min\left(x - \boldsymbol{q}_1, x - \boldsymbol{q}_2, \|x - (\lambda q_1 + (1 - \lambda)\, q_2)\|\right) \tag{5}$$

where x represents the input vector, $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are the two endpoints of the super sausage, respectively. $\|\cdot\|$ indicates a distance metric, which is the Euclidean distance. λ represents the projected length of $\boldsymbol{q}_1 x$ on the vector $\overrightarrow{\boldsymbol{q}_1\boldsymbol{q}_2}$, which is expressed as follows:

$$\lambda = \begin{cases} 1, & \boldsymbol{k}\left(\boldsymbol{q}_2 - \boldsymbol{q}_1\right) < 0 \\ \|\boldsymbol{k}\|, & \frac{\|\boldsymbol{k}\|}{\|\boldsymbol{q}_2\boldsymbol{q}_1\|} < 1, \quad \boldsymbol{k}\left(\boldsymbol{q}_2 - \boldsymbol{q}_1\right) \geqslant 0 \\ 0, & \frac{\|\boldsymbol{k}\|}{\|\boldsymbol{q}_2\boldsymbol{q}_1\|} \geqslant 1, \quad \boldsymbol{k}\left(\boldsymbol{q}_2 - \boldsymbol{q}_1\right) > 0 \end{cases} \tag{6}$$

where $\boldsymbol{k}$ refers to the projection vector of $\boldsymbol{q}_1 x$ on the vector $\overrightarrow{\boldsymbol{q}_1\boldsymbol{q}_2}$, calculated as follows:

$$\boldsymbol{k} = \frac{\left(x - \boldsymbol{q}_1\right)\left(\boldsymbol{q}_2 - \boldsymbol{q}_1\right)}{\|\boldsymbol{q}_2 - \boldsymbol{q}_1\|} \tag{7}$$

where $d = \|x - (\mathbf{q}_1 + (1 - \lambda)\mathbf{q}_2)\|$ is the distance from the input x to the vector $\overrightarrow{\boldsymbol{q}_1\boldsymbol{q}_2}$. If $\frac{d}{r} > 1$, the corresponding sample is outside the hypersausage geometry and can be considered to be a negative sample. Otherwise, the corresponding sample is inside the hypersausage geometry and functions as a positive sample, in which d indicates the distance between the sample point and the vector. Besides, the feature points with the same shape are consistent with the same distance calculation method.

Step 2: Compute the measured output y of each sausage, which can be determined

$$y = \frac{s}{\|s\|} \cdot \exp\left(-\frac{s^2}{2\gamma^2}\right) \tag{8}$$

where $y \in [0, 1]$ denotes the predicted i neuron output, r expresses the radius of the hypersausage metric. The first term $\frac{s}{\|s\|}$ indicates the location of the lower capsule in relation to the upper capsule. The second term $\exp(-\frac{s^2}{2\gamma^2})$ represents the likelihood of the characteristic indicated by the upper Capsule's presence.

IV. Experiments

The proposed method's feasibility is confirmed through detailed and comprehensive experiments on two public remote sensing detection datasets. This section describes the datasets, evaluation metrics, and training details.

A. Dataset Description

The DIOR dataset [53] refers to a large, diverse, and object detection dataset that is available publicly in the earth observation community. It exhibits the characteristics as follows. 1) The dataset is large-scale in terms of the object types, object instance numbers, and total number of images. It comprises 23 463 images and 192 472 instances including 20 object classes. 2) The dataset contains a large range of object sizes, including different sizes among similar targets and different sizes among different types of objects. 3) The dataset also has changed in the appearance of objects due to various imaging conditions, weather conditions, seasons, as well as image qualities. 4) The dataset contains high-class intersimilarity and diversity within classes. To ensure that similar distributions of test data and training-validation (train-val) data, the training set selects 11 725 remote sensing images (i.e., 50% of the dataset). In addition, the test set contains the remaining 11 738 images. The train-val data comprise two parts, including the validation (Val) set and training (train) set.

The HRRSD dataset [54] is a public dataset for multiclass object detection in optical remote sensing images, including 13 types of typical man-made objects including ships (SP), bridges (BG), ground runways (GTF), and storage tanks (ST). Target. In the DOIR-R dataset, there are 21 761 pictures and 55 740 instances, and the number of images accounts for 0.25, 0.25, and 0.5 in the training, verification, and testing set separately.

B. Evaluation Metrics

The mean average precision (mAP) acts as the evaluation metric for the proposed method, similar to most of object detection methods. The definition of mAP is

$$mAP = \frac{1}{K}\sum_{n=1}^{K}\int\left(P_n\left(R_n\right)\right)dRn \tag{9}$$

where R_n refers to the recall for a given class n, $P_n(R_n)$represents the precision since R_n indicates the recall of this class. K represents the classes' total number. The precision and recall are written as follows:

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$P = \frac{TP}{TP + FP} \tag{11}$$

where TP, FN, and FP suggest the numbers of true positives, false negatives, and false positives, respectively.

C. Implementation Details

The ResNet50 [6] model pretrained on ImageNet [7] becomes the backbone network in the current work. We train all models on the training set. Later, the models are tested on the testing set. The subimages are resized to 800 × 800 pixels at the training and validation stages. With 1 image per GPU, the models are trained on 4 GPUs. The proposed model is trained on the training set for 50 epochs with the AdamW optimizer [55]. The initial learning rate is determined to be 10^{-5} regarding the backbone and 10^{-4}

TABLE I
MEAN AVERAGE PRECISION (mAP) SCORES OF VARIOUS METHODS, IN WHICH BC DENOTES THE BASKETBALL COURT, ESA REPRESENTS THE EXPRESSWAY SERVICE AREA, ETS SUGGESTS THE EXPRESSWAY TOLL STATION, AND GTF DENOTES THE GROUND TRACK FIELD; IN ADDITION, THE ENTRIES WITH THE BEST APs FOR EACH OBJECT CATEGORY ARE SHOWN TO BE BOLD FACED

Categories	Faster RCNN	**SSD**	RFBNet	RetainNet	YOLOv3	YOLOv3 -ASFF	EifficentDet	FRPNet	CSFF	CF2PN	Ours
Airplane	36.39	67.42	77.97	68.63	53.92	91.86	71.18	64.5	57.2	78.32	**80.33**
Airport	57.34	61.33	76.89	70.95	57.86	71.11	69.51	**82.6**	79.6	78.29	74.29
Baseball Field	62.44	68.23	74.15	73.00	64.84	74.54	67.97	**77.7**	70.1	76.48	71.67
BC	75.63	85.94	88.71	89.42	79.19	89.34	86.72	81.7	87.4	88.4	**89.11**
Bridge	18.36	25.63	34.35	32.49	30.08	33.33	38.47	**47.1**	46.1	37	43.87
Chimney	72.60	76.63	**78.84**	76.79	68.52	75.77	76.06	69.6	76.6	70.95	78.56
Dam	42.69	51.00	58.39	**65.74**	47.83	43.13	58.18	50.6	62.7	59.9	60.24
ESA	52.30	59.77	72.19	**85.71**	56.63	58.12	58.27	80.0	82.6	71.23	60.27
ETS	43.90	46.11	54.34	55.13	47.13	57.28	55.15	71.7	**73.2**	51.15	62.62
Golf Course	63.49	73.39	80.41	**82.11**	59.30	73.81	73.26	81.3	78.2	75.55	75.39
GTF	48.68	63.06	71.48	80.82	54.38	46.42	69.28	77.4	**81.6**	77.14	74.55
Harbor	29.22	51.72	61.26	54.31	48.28	57.11	59.00	**78.7**	50.7	56.75	61.76
Overpass	43.27	50.89	57.15	55.26	46.95	56.21	54.72	**82.4**	59.5	58.65	57.28
Ship	12.62	54.69	76.30	43.58	76.95	75.11	85.42	62.9	73.3	76.06	**89.48**
Stadium	38.15	67.45	79.75	**81.62**	40.54	37.92	42.49	72.6	63.4	70.61	54.21
Storage Tank	17.16	39.76	51.42	36.26	54.18	**77.33**	67.34	67.6	58.5	55.52	74.52
Tennis Court	65.86	82.41	86.81	86.70	76.58	**89.21**	81.32	81.2	85.9	88.84	85.44
Train Station	41.47	43.60	61.57	48.07	42.56	49.51	48.14	**65.2**	61.9	50.83	57.70
Vehicle	11.39	25.89	35.50	20.11	33.84	52.74	45.86	52.7	42.9	36.89	**56.05**
Windmill	41.33	63.26	74.61	67.45	67.10	82.00	75.21	**89.1**	86.9	86.36	81.63
mAP	43.71	57.91	67.61	63.71	55.33	64.59	64.18	71.8	68.0	67.25	**69.45**

TABLE II
COMPARISON WITH REPRESENTATIVE DETECTORS ON HRRSD DATASET; BEST RESULTS ARE SHOWN IN BOLD

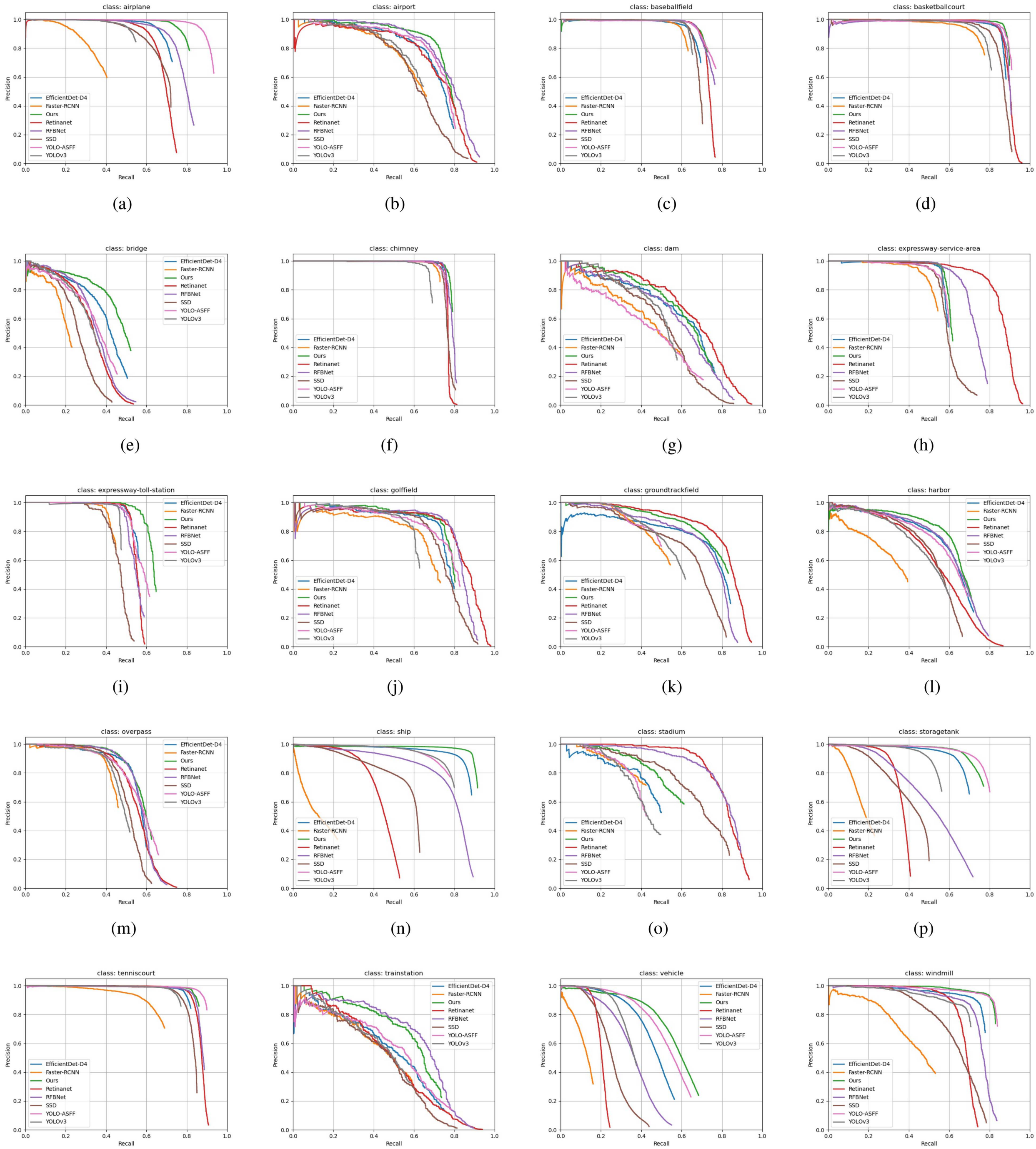
Methods	YOLOv3	FCOS	FRCNN	RetinaNet	HIE-Det	S2BDet	MSE-Net	GLFPN	Ours
mAP	71.80	82.26	83.10	83.53	88.70	85.20	86.50	84.18	**88.72**

regarding the rest. The learning rate declines by 0.1 factor at the 40th epoch. Besides, the weight decay is defined as 10^{-4}, and the dropout rate is defined as 0. The feed-forward network hidden dimension is 1024, the attention feature channel reaches 256, and the attention head is 8. Moreover, a set of learned points are used to be the anchor points by default. Additionally, the number of encoder layers and decoder layers is 6, which is similar to DETR. Data augmentations and tradeoff hyperparameters in detection loss are the same as DETR.

D. Comparison with State-of-the-Art Methods

Results on DIOR: In the present study, we compared the experimental findings to various typical deep learning-based approaches that are often adopted for object recognition in natural images and remote sensing images to quantitatively assess the proposed method's detection performance. The chosen methods contain one-stage and two-stage methods. To be specific, the compared methods contain Faster RCNN [28], SSD [29], RFBNet [56], RetinaNet [47], YOLOv3 [57], YOLOv3-ASFF [58], EifficentDet [59], FRPNet [60], CSFF [61], and CF2PN[62]. In order to make fair comparisons, this study maintains all the experimental settings the same as those presented in the corresponding papers. For SSD, RFBNet, FRPNet, CF2PN, and Faster RCNN, we use VGG16 as their backbone networks. Regarding YOLOv3 and YOLOv3-ASFF, their backbone networks use the Darknet-53 framework. For FRPNet and CSFF, they use ResNet-101 as the backbone network, respectively. RetainNet uses ResNet-50 as the backbone network. For EifficientDet, we use EifficienNet-B4 as the backbone network.

Table I presents the APs of all methods of various types, and Fig. 4 lists the precision-recall (PR) curves of various types. Based on Table I, the proposed method is significantly better when compared with the other methods and exhibits the highest mAP of 69.45%. According to Table I, the mAP value of Faster RCNN is the lowest, while the AP value for vehicles is merely 11.39%. The reason is that Faster RCNN only adopts a single-layer feature map for performing regression positioning on the target, such that it cannot express multiscale information and exhibits poor detection results for small and large targets. The mAP values of SSD, RFBNet, RetinaNet, and YOLOv3 could all reach over 57% due to that they all employ multiscale feature layers for predicting and obtaining perceptual fields of varying sizes by various scale feature layers, thereby enhancing the accuracy of detection. In addition, the performance of SSD is lower than that of the above algorithm. This is because the SSD algorithm neither uses the surrounding context information to assist small target detection nor does it incorporate the feature fusion mechanism of the FPN. Compared with SSD, RFBNet enhances the diversity of features through a multilayer receptive field fusion mechanism and uses the surrounding context information to assist small target detection. Moreover, the experiment proves that the RFB module enhances the multiscale expression ability. However, RFBNet lacks a multiscale feature fusion



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

(q) (r) (s) (t)

Fig. 4. Precision-recall (PR) curves of different methods. (a) Airplane. (b) Airport. (c) Baseball Field. (d) Basketball Court. (e) Bridge. (f) Chimney. (g) Dam. (h) Expressway Service Area. (i) Expressway Toll Station. (j) Golf Field. (k) Ground Track Field. (l) Harbor. (m) Overpass. (n) Ship. (o) Stadium. (p) Storage Tank. (q) Tennis Court. (r) Train Station. (s) Vehicle. (t) Windmill.

module and does not consider the problem of scale imbalance, which makes its performance inferior to that of our proposed algorithm.

Compared with the Faster RCNN and SSD methods, RetainNet and YOLOv3 use a feature pyramid structure to improve their multiscale expression capabilities, but the feature fusion mechanism they use still does not consider the semantic differences of features at varying scales and neglects the semantic information of other feature layers. These shortcomings cause their performance to be inferior to that of our algorithm. Based on experiments that our method performs better than YOLOv3 and RetainNet by 5.74% and 13.9%, which verifies the effectiveness of the proposed multiscale feature adaptive fusion strategy.

According to the experimental findings, the detection results of YOLOv3-ASFF and EifficentDet are better than those of RetainNet and YOLOv3, which use the classic feature pyramid. The reason is that the first three detection networks superimpose

TABLE III
Ablation Study on the Components of the Scale-Aware Pyramid Network

Baseline	Multilevel feature fusion	Capsule reasoning	Sausage metrics	mAP%
✓				63.71
✓	✓			65.31
✓	✓	✓		68.51
✓	✓	✓	✓	69.45

and combine the multiscale features obtained by the basic network to obtain new multilevel and multiscale features to achieve the advantages of the strong-aggregation shallow-information positioning ability and deep information classification ability. The advantage of this structure is that it can overcome the weakness where each feature map in the classic feature pyramid structure is mostly composed of single-level features, which causes poor FPN feature expression results.

Although the above methods show promising results, they overlook the issue of scale imbalance, where the number of small targets significantly exceeds that of large targets. Consequently, the weights assigned to small-scale targets during training are weaker than those of large-scale targets. Notably, our proposed approach showcases stronger small-scale target detection ability relative to the above three methods, thanks to the following three strategies. Nonetheless, to tackle the problem of scale imbalance, we prioritize employing specialized measures in our approach. Our experimental results demonstrate the superiority of the proposed method in detecting small-scale targets compared to the above three methods. 1) It adopted the long-distance dependence modeling of the self-attention layer for replacing the locality modeling of the convolutional layer, which can show the global interaction between heterogeneously distributed objects, and show significant advantages at distinguishing their types and locations from the clutter background. 2) Capsule-inference module can explore the object entity information, and utilize the bidirectional attention routing for forward information delivery and backward information feedback.

Results on HRRSD: With the purpose of further confirming the effectiveness of the proposed method, this study compares the performance of the proposed model with the classic detection model on the HRRSD dataset. Specifically, the compared methods include YOLOv3, FCOS, FRCNN, RetinaNet, HIE-Det [63], S2BDet [64], MSE-Net [65], and GLFPN [66]. Table III presents the findings of the performance comparison. Based on Table II, the proposed method achieves 88.72% mAP on the HRRSD dataset. Our method still achieved good results, which further verifies the superiority of capsule reasoning and sausage metrics.

E. Ablation Study

This section examines the impact of each module of the proposed network on performance. The ablation experiments focus on elucidating the two parts of Capsule reasoning and Sausage Metrics. To demonstrate the feasibility of the proposed approach, we train and conduct experiments on the DIOR dataset. Table III presents several comparisons evaluating the contributions of each module. First, we evaluate some components' contribution to the baseline recognizer of this study, which provides a reference. Overall, the techniques enhance accuracy, resulting in a final baseline mAP score of 69.45%.

1) Feature Fusion With Multiple Levels: We introduce a feature fusion method with multiple levels to learn balanced semantic features by integrating and refining features at all FPN levels. This approach fully utilizes multi-scale feature information, without increasing the model training burden. The experimental results show that after introducing the Capsule Reasoning module, the map value increases by 1.6%, compared to the baseline.

2) Capsule Reasoning: The reasoning capsule module employs capsule construction and attention routing to realize target type and location information reasoning and avoids the limitations of describing the target's relative position when using the FFN model directly. This model's autonomous reasoning ability enhances target detection performance. The experimental results show that after introducing the Capsule Reasoning Module, the map value increases by 4.8% compared to the baseline. This improvement stems from constructing capsule entities that achieve a comprehensive internal target representation and obtaining a more effective inference method for the target from capsule entities through the routing information transmission method between them.

3) Sausage Metrics: A hypersausage metric model is introduced in the capsule inference module to obtain the probability of the respective attribute of each target capsule appearing in the target. This metric function exhibits strong nonlinear ability and can effectively describe the underlying capsules (features). Besides, the mapping relationship between high-level capsules (types and positions) enhances the expressiveness of features. Table III presents the comparative findings of the ablation experiments based on the sausage metric module in the CI_DETR model. As depicted in Table III, when the squash activation function in the capsule is replaced with the hypersausage metric model in this study, the detection accuracy of the model has been significantly increased.

F. Experimental Results and Analysis

To reveal the performance of the proposed method visually, this study explains it from two perspectives, qualitative and quantitative. As shown in the qualitative analysis, this study visualized the detection results in the DIOR dataset, and the results are presented in Fig. 5. As depicted in Fig. 5, the proposed method has better performance under varying scale targets and different backgrounds; CI_DETR makes a good performance in detecting not only small, dense objects (e.g., small ships, oil tanks, vehicles, and airplanes), but also large objects (e.g., bridges, basketball courts, ground track fields, and overpasses). In the DIOR dataset, vehicles are small targets that have a large

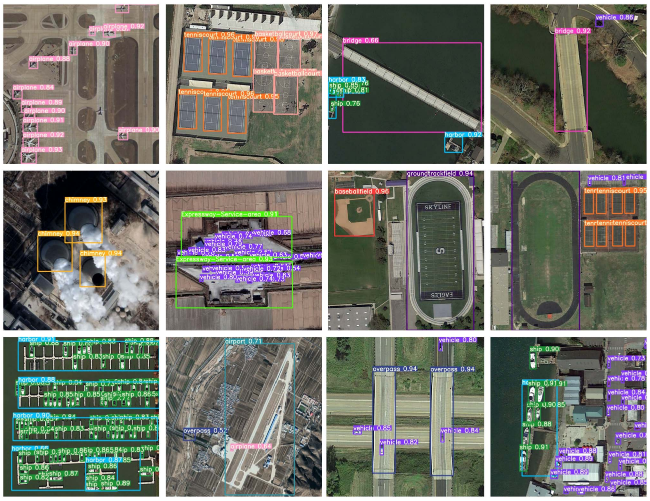


Fig. 5. Detection results of the proposed methods.

sample size, and their scene complexity is higher than that of other targets. We have enhanced our method by incorporating the Multi-level Feature Fusion module within the feature pyramid structure. This allows for better extraction of contextual information surrounding small targets, thereby improving the overall mining ability of the system. In addition, we have utilized a capsule inference model that constructs the capsule entity, resulting in a more comprehensive internal representation of the target. By leveraging the routing information transmission method between the capsule entities, our system is able to more effectively infer information about the target.

V. Conclusion

In this paper, we propose a multi-scale feature aggregation module to improve the detection ability and fusion of small-scale features in multi-scale targets. Firstly, we incorporate a capsule reasoning model for entity mining, where the object category and location are predicted using attention routing. Additionally, we employ a sausage metric model, which has a strong nonlinear mapping ability and can effectively predict capsule entities to describe the mapping relationship between features and type labels. This model's detection performance is further enhanced by predicting the probability of the target's existence. Experimental results on a public remote sensing dataset demonstrate that our proposed method outperforms other object detection methods like RetainNet and EfficientDet.

References

[1] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Contextual classification of lidar data and building object detection in urban areas," *ISPRS J. Photogrammetry Remote Sens.*, vol. 87, pp. 152–165, 2014.

[2] Z. Chen, T. Zhang, and C. Ouyang, "End-to-end airplane detection using transfer learning in remote sensing images," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 139.

[3] L. Tang, W. Tang, X. Qu, Y. Han, W. Wang, and B. Zhao, "A scale-aware pyramid network for multi-scale object detection in SAR images," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 973.

[4] C. Deng, D. Jing, Y. Han, S. Wang, and H. Wang, "FAR-Net: Fast anchor refining for arbitrary-oriented object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6505805.

[5] Y. Han, H. Liu, Y. Wang, and C. Liu, "A comprehensive review for typical applications based upon unmanned aerial vehicle platform," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9654–9666 2022.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778. .

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[8] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-aware anti-drift object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4075–4086, Aug. 2019.

[9] Y. Han, C. Deng, B. Zhao, and B. Zhao, "Spatial-temporal context-aware tracking," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 500–504, Mar. 2019.

[10] C. Deng, Y. Han, and B. Zhao, "High-performance visual tracking with extreme learning machine framework," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2781–2792, Jun. 2020.

[11] B. Zhao, B. Zhao, L. Tang, Y. Han, and W. Wang, "Deep spatial-temporal joint feature representation for video object detection," *Sensors*, vol. 18, no. 3, 2018, Art. no. 774.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, 2014.

[14] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[16] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 3–22, 2018.

[17] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–868, Jun. 2019.

[18] Z. Zheng et al., "Hynet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 1–14, 2020.

[19] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 4096–4105.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[21] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021.

[22] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A. spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 4409512.

[23] X. Xu et al., "An improved swin transformer-based model for remote sensing object detection and instance segmentation," *Remote Sens.*, vol. 13, no. 23, 2021, Art. no. 4779.

[24] Q. Li, Y. Chen, and Y. Zeng, "Transformer with transfer CNN for remote-sensing-image object detection," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 984.

[25] H. Sebastian Seung, "Learning continuous attractors in recurrent networks," in *Proc. 1997 Conf. Adv. Neural Inf. Process. Syst. 10 (NIPS '97)*, Cambridge, MA, USA: MIT Press, 1998, pp. 654–660.

[26] X. Ning, Y. Wang, W. Tian, L. Liu, and W. Cai, "A biomimetic covering learning method based on principle of homology continuity," *ASP Trans. Pattern Recognit. Intell. Syst.*, vol. 1, no. 1, pp. 9–16, 2021.

[27] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 3859–3869, 2017.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[29] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[31] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[32] H. Law and J. Deng, “Cornernet: Detecting objects as paired keypoints,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[33] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “Centernet: Keypoint triplets for object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578.

[34] X. Zhou, J. Zhuo, and P. Krahenbuhl, “Bottom-up object detection by grouping extreme and center points,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 850–859.

[35] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS’17)*, Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[36] Z. Dai, B. Cai, Y. Lin, and J. Chen, “UP-DETR: Unsupervised pre-training for object detection with transformers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1601–1610.

[37] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[38] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, “ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art no. 5614914.

[39] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, “Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 294–308, 2020.

[40] L. Liu, Z. Pan, and B. Lei, “Learning a rotation invariant detector with rotatable bounding box,” 2017, *arXiv:1711.09405*.

[41] H. Shi, Z. Fang, Y. Wang, and L. Chen, “An adaptive sample assignment strategy based on feature enhancement for ship detection in SAR images,” *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2238.

[42] H. Shi, B. Chai, Y. Wang, and L. Chen, “A local-sparse-information-aggregation transformer with explicit contour guidance for SAR ship detection,” *Remote Sens.*, vol. 14, no. 20, 2022, Art. no. 5247.

[43] H. Shi, C. He, J. Li, L. Chen, and Y. Wang, “An improved anchor-free SAR ship detection algorithm based on brain-inspired attention mechanism,” *Front. Neurosci.*, vol. 16, 2022, Art. no. 1074706.

[44] B. Zhao, Y. Wu, X. Guan, L. Gao, and B. Zhang, “An improved aggregated-mosaic method for the sparse object detection of remote sensing imagery,” *Remote Sens.*, vol. 13, no. 13, 2021, Art. no. 2602.

[45] B. Zhao, Q. Wang, Y. Wu, Q. Cao, and Q. Ran, “Target detection model distillation using feature transition and label registration for remote sensing imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5416–5426, 2022.

[46] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu, and Z. Li, “Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5786–5795, 2021.

[47] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[48] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, “Libra R-CNN: Towards balanced learning for object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821–830.

[49] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[50] R. Al-Rfou, D. Choe, N. Constant, M. Guo, and L. Jones, “Character-level language modeling with deeper self-attention,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 3159–3166.

[51] E. Xi, S. Bing, and Y. Jin, “Capsule network performance on complex data,” 2017, *arXiv:1712.03480*.

[52] X. Ning et al., “BDARS_CapsNet: Bi-directional attention routing sausage capsule network,” *IEEE Access*, vol. 8, pp. 59059–59068, 2020.

[53] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[54] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, “Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.

[55] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Representations*, 2018.

[56] S. Liu and D. Huang et al., “Receptive field block net for accurate and fast object detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 385–400.

[57] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” 2018, *arXiv:1804.02767*.

[58] S. Liu, D. Huang, and Y. Wang, “Learning spatial fusion for single-shot object detection,” 2019, *arXiv:1911.09516*.

[59] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.,* 2019, pp. 6105–6114.

[60] J. Wang, Y. Wang, Y. Wu, K. Zhang, and Q. Wang, “FRPNet: A feature-reflowing pyramid network for object detection of remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art no. 8004405.

[61] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, “Cross-scale feature fusion for object detection in optical remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2021.

[62] W. Huang, G. Li, Q. Chen, M. Ju, and J. Qu, “CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection,” *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 847.

[63] Y. Zhang and Y. Yuan, “Hierarchical information enhancing detector for remotely sensed object detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2022, Art no. 6000405.

[64] Y. Liu, Q. Li, Y. Yuan, and Q. Wang, “Single-shot balanced detector for geospatial object detection,” in *Proc. IEEE Int. Conf. Acoust.,Speech Signal Process.*, 2022, pp. 2529–2533.

[65] H. Lv, W. Qian, T. Chen, H. Yang, and X. Zhou, “Multi-scale feature adaptive fusion for object detection in optical remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art no. 6511005.

[66] N. Liu, T. Celik, and H.-C. Li, “Gated ladder-shaped feature pyramid network for object detection in optical remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021, Art no. 6001505.

Yingchao Han is currently a Doctor with the Harbin Institute of Technology, Harbin, China, and a Senior Engineer. He has authored or coauthored more than 30 papers in journals and international conferences. He has applied for more than 20 patents for inventions. His research interests include aircraft system networking communication, and design.

Weixiao Meng (Senior Member, IEEE) received the bachelor’s degree in engineering, major in electronic instrumentation and measurement technology, from the Department of Electrical Engineering, Harbin Institute of Technology, Harbin, China, in 1990, the master’s degree in engineering, major in communication engineering, from the Department of Radio Engineering, Harbin Institute of Technology, in 1995, and the doctoral degree in engineering, major in communication and information systems, from the Department of Electronics and Communication Engineering, Harbin Institute of Technology, in 2000.

He is currently a Full Professor with the School of Electronics and Information Engineering, HIT. He has authored or coauthored four books and over 300 papers in journals and international conferences. His research interests include broadband wireless communications, space-air-ground integrated networks, and wireless localization technologies.

Dr. Meng is the Chair of the IEEE Communications Society Harbin Chapter, a Fellow of the China Institute of Electronics, and a Senior Member of the IEEE ComSoc and the China Institute of Communication. He acted as leading TPC Cochair of ChinaCom2011 and ChinaCom2016, leading Services and Applications track Co-chair of IEEE WCNC2013, Awards Cochair of IEEE ICC2015 and Wireless Networking Symposia Cochair of IEEE Globecom2015, AHSN Symposia Cochair of IEEE Globecom2018, leading Workshop Cochair of IEEE ICC2019 and IEEE ICNC2020, AHSN Symposia Cochair of IEEE ICC2020. In 2005, he was honored provincial excellent returnee and selected into New Century Excellent Talents (NCET) plan by the Ministry of Education (MOE), China in 2008, and the Distinguished Academic Leader of Harbin. Under his leading, Harbin Chapter won IEEE ComSoc Chapter of the Year Award, the Asia Pacific Region Chapter Achievement Award, and personal Member and Global Activities Contribution Award in 2018. In 2021, he won the Best Paper Award of IEEE System Journal.

Wei Tang (Member, IEEE) received the B.Eng. degree in electronic information science and technology from the School of Electronic Science and Engineering, Jilin University, Changchun, China, in 2016, and the Ph.D. degree in communication and information engineering from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 2022.

His research interest include computer vision and deep learning, especially on object detection, object recognition, and remote sensing.