Dual Network Structure With Interweaved Global-Local Feature Hierarchy for Transformer-Based Object Detection in Remote Sensing Image

Jingqian Xue, Da He, and Qian Shi

Abstract—Frequent and accurate object detection based on remote sensing images is an encouraging approach for monitoring dynamic of the interest object on earth surface. Transformer-based object detection was recently developed to cope with the tradeoff dilemma between large computation load and accuracy sacrifice encountered by region-proposal-based and regression-based object detection, and its self-attention mechanism can provide a global understanding that has potential ability for reasoning the location relationship within sparsely heterogeneous distributed geospatial objects. However, transformer-based object detection is essentially weak at modeling local feature hierarchy to compensate for the large scale variation of geospatial object, and it is extremely difficult to train due to the lack of inductive bias, resulting in a slow convergence. To overcome the problem, this article proposed a Dual network structure with Interweaved Global-local feature hierarchy based on the TRansformer architecture (DIAG-TR), to alleviate the incompatibility of global and local feature form, and hierarchically embed the local features into global representations. Besides, a learnable anchor box is incorporated into the positional query in the decoder part to provide a spatial prior, which can accelerate convergence. The proposed DIAG-TR is validated on the widely used optical remote sensing image DIOR dataset, and the results demonstrate that the global-local feature hierarchy contributes 3.4% mean average precision compared to the original transformer-based method, and the convergence time is shortened by 2.5-fold. State-of-the-art methods are also participated as benchmark for comparison, and DIAG-TR outperforms baseline method by 8.9%, which proves that DIAG-TR has great potential in earth observation community.

Index Terms—Global-local feature hierarchy, object detection, optical remote sensing image, self-attention mechanism, transformer.

I. INTRODUCTION

With the increasing of spatial resolution of remote sensing images, frequent and accurate recognition of the interest geospatial object from earth observation becoming crucial for wide applications, such as illegal construction in urban planning [1], [2], military reconnaissance [3], and airplanes and vehicle monitoring for traffic controlling [4], [5]. Object detection is one of the most major techniques to undertake the two main tasks, i.e., automatic recognition of the object and localization of the precise geolocation [6], [7].

In the past decades, various of deep-learning-based object detection methods have been proposed in the computer vision community, and achieved impressive performance. Generally, candidate bounding box/anchor of the object should be first selected from the global scene, and the convolutional neural network (CNN) features of each candidate are extracted, which are then fed to a classifier for determining if the bounding box really contains the object, and recognizing the categories of the object [8]. In this way, the location and category of the object are accomplished.

However, on one hand, different from the natural images that are usually taken from profile aspect consisting of clear scene with obvious target, the earth observation images are taken by satellite/aerial photography from top-view aspect, the number, scale, rotation, and geometric distortion of the geospatial objects are various, which results in an exponential growth of the possible candidate bounding box to search [9], [10], and [11]. On the other hand, the geospatial object usually heterogeneously distributed and mixed with clutter background, which makes the traditional convolutional manner with limited receptive field difficult to understand the global context of the geographic scene [12], [13]. Current CNN commonly adopted stacked convolutional layer and pooling layer to expand the receptive field through downsampling, however, the geospatial object is usually much smaller and indistinctive compared to the whole scene, which will result in the miss of small object [14]. Therefore, frequent and accurate geospatial object detection from earth observation remains a challenge task.

Recently, transformer-based object detection method was developed [15], [16], and [17]. It adopted the long-distance dependency modeling of the self-attention layer to replace the locality...
modeling of the convolutional layer, which can represent the global interaction between heterogeneously distributed objects, [18], and has great advantages at distinguishing their categories and locations from the clutter background [11], [19], [20]. Besides, different from the traditional object detection manner that use pair-wise prediction and ground truth defined by the region-proposal/anchors for training, transformer-based method reformulate the detection problem to a disorder set prediction and matching paradigm, which can automatically match the prediction (category and bounding box) with its ground truth, and relieve the need for hand-designed region-proposal/anchors [15]. Therefore, it has great potential for applying geospatial object detection in earth observation images.

However, the local features represented by shallow CNN, such as texture, edge, shape, are important for small-scale object recognition, which are deteriorated by the stacked self-attention layer in transformer-based method, resulting in a relatively poor performance on small-scale geospatial object [19]. Although some works have attempted to fuse local and global feature to alleviate this problem, the local and global information is separately processed, in which the subsequent self-attention layer may overwrite the local features. Besides, transformer network is extremely difficult to train [17], due to the lack of inductive bias, resulting in a slow convergence.

Aimed at alleviating these bottlenecks, a Dual network structure with Interweaved Global-local feature hierarchy upon the TRansformer-based object detection (DIAG-TR) is proposed in this article. In DIAG-TR, a global-local feature interweaving (GLFI) module is designed in the encoder of the transformer architecture, based on the interweaved convolutional branch and self-attention branch, to adaptively and hierarchically embed the local features into global representations, and compensate for the scale variation of geospatial object. Considering that the CNN features (two-dimension map with coordinate) are incompatible with the global features (sequential token without coordinate), a feature resampling process is designed in GLFI module to bidirectionally transforms global and local feature. Besides, an anchor box coordinate is incorporated into the object query as inductive bias in the decoder part, to explicitly guide the learning of network and accelerate the convergence.

The main contributions of this article are as follows.

1. A dual network structure with convolutional branch and self-attention branch is designed to hierarchically embed the local features into global representations, to consider the large scale variation of geospatial object. Besides, a feature reconstruction mechanism is designed to mitigate the incompatibility of the CNN feature and global feature representation for a better incorporation.

2. An anchor box coordinate is adopted to be embedded in the decoder object query in each decoder layer, to provide inductive bias for learning the spatial locations to accelerate convergence of the network model.

3. The proposed method is validated on the largest object detection dataset DIOR [7], which outperforms the state-of-the-art Faster-RCNN-FPN and original transformer-based method by 8.9% and 3.4% mean average precision, respectively.

The rest of this article is organized as follows. Foundation and literature of the deep-learning-based object detection method are reviewed in Section II. The detail construction of the proposed DIAG-TR is developed in Section III. Section IV provides a validation and ablation experiment, and makes a comparison with the state-of-the-art method. Finally, Section V concludes this article.

II. RELATED WORK

Numerous deep-learning-based approaches for object detection of natural scene images in visual community have been explored to detect various geospatial objects in the earth observation community, and the contemporary object detection methods can be summarized into two categories according to whether involving region-proposal, i.e., region-proposal based two-stage method and regression-based one stage method.

A. Region-Proposal-Based Methods

Region-proposal-based CNN method (R-CNN) is one of the most prevalent hotspots in object detection field [8], [14], [21], [22], [23], and [24]. This method first generates candidate region proposals for the possible objects in the input image, and then use CNN to extract features of each region proposal, which are finally processed by classifiers to recognize the category of the object, and localize the precise minimum bounding rectangle box of the object. Considering the time-consuming during the duplicated feature extraction of CNN for almost 2000 region-proposals, Fast R-CNN [22] was developed to implement CNN feature extraction only once for the whole input image, and identify the features of each region-proposals on the whole feature map through corresponding projection. However, the generation of region-proposals is still based on hand-designed selective search method, which is kept separate from the subsequent process and will result in computational load. To this end, Faster R-CNN [23] is proposed to replace the traditional selective search method by a region proposal network (RPN) embedded in the CNN feature extraction network, which can learn the candidate anchors instead of region-proposals, and it can be trained end-to-end, which further reduces time consumption.

Most of the current object detection methods in earth observation community are based on Faster R-CNN, and further ameliorations have been made to account for the characteristic of the geospatial object, such as rotation, geometric variation, scale variation, and context. Considering the orientation of the rotated geospatial object, Li et al. [25] explored the feasibility of multangle anchors in RPN to precise capture the object; Cheng et al. [26], [27] proposed to embed a new rotation-invariant CNN (RICNN) model into R-CNN framework for object detection. Considering the complex context of geospatial object, Zhong et al. [28] utilized a position-sensitive balancing method to handle the translation-variance in object detection and precisely localize the bounding box of the object, and Long et al. [29] propose an unsupervised score-based bounding box regression algorithm to refine the region of the bounding box. Considering the scale variation of different geospatial object [30], [31], Zhang et al. [32] adopted feature-pyramid network (FPN)
to embed into Faster R-CNN to hierarchically extract the multiscale features and increase the performance on small-scale geospatial object; Liu et al. [33] proposed a path aggregation network (PANet) to fuse the low-level location features with the topmost features by bottom-up path augmentation to increase the small-scale object performance. Considering the geometric variation of geospatial object, Xu et al. [34] and Dai et al. [35] proposed a deformable R-FCN to adaptively model the irregular geometric of the object. Other works are also dedicated to the specific geospatial object, such as airplane [3], vehicle [36], and ship [37]. However, the size, aspect ratio, or angles of the anchors are also hand-designed with discrete value, which means that a large number of anchors may be needed to approach to the real bounding of the object, and consequently increase the computational burden.

B. Regression-Based Methods

In order to realize real-time detection, one-stage regression-based methods are developed, which can directly predict both object categories and bounding box without the need for region-proposals and, thus, largely promotes the computational efficiency. Numerous one-stage methods have also been deployed into the remote sensing images for geospatial objection detection, benefit to their advantage of real-time detection. Typical model includes You Only Look Once (YOLO) [37], [38], and [39], single shot multibox detector (SSD) [40], [41], and RetinaNet[42]. For example, Tan et al. [36] used a regression-based object detector to detect vehicle targets with arbitrary-oriented bounding boxes; Liu et al. [43] adopted the rotatable bounding box in SSD framework to estimate the orientation angle of the object, and further designed an arbitrary-oriented YOLOv2 to detect ships [44]; Wang et al. [45] adopted RetinaNet to detect ship on the Gaofen-3 imagery. However, a tradeoff always exists between computational efficiency and detection performance [7], [46], and regression-based method may sacrifice the detection performance.

C. Transformer-Based Method

Recently, transformer architecture was introduced to the object detection community [15], which reframes the object detection problem to a direct set prediction problem, without any hand-engineering region-proposal, anchor, or window operations, and still reaches the state-of-the-art accuracy. Furthermore, benefit to the global representation of transformer compared to CNN, transformer-based detector can infer the long-distance relationship between object and global image context, and better handle the complex cluttered background in earth observation images. Carion et al. [15], first, applied transformer architecture (DETR) to solve the object detection problem, Beal et al. [47] adopted transformer as backbone for the traditional R-CNN (ViT-FRCCN), Sun et al. [48] discussed the several factors contributing to the slow convergence of DETR, Xu et al. [49] proposed a local-perception Swin Transformer backbone for geospatial object detection, Ma et al. [50] designed an oriented DETR to compensate for rotated object. Considering the slow convergence of DETR, Zhu et al. [16] introduced the idea of deformable convolution into transformer architecture to only focus on a small set of key sampling points and accelerate convergence, and Liu et al. [17] provided a clear understanding of the content query and anchor-box-based coordinate query in DETR to further accelerate convergence. Other works, such as DN-DETR [51], DINO [52], are also dedicated for better convergence of DETR. Therefore, transformer-based object detection is a promising way to solve the tradeoff dilemma of CNN-based detection.

However, the current transformer-based detection is seldom introduced into the earth observation community due to the following reasons: 1) Transformer-based detector is weak at learning the feature hierarchy with low-level localization information and difficult to detect small-scale object; 2) without inductive bias about the target object, transformer network is extremely difficult to train, resulting in a slow convergence.

III. DUAL NETWORK STRUCTURE WITH INTERWEAVED GLOBAL-LOCAL FEATURE HIERARCHY FOR TRANSFORMER-BASED OBJECT DETECTION

In order to address the shortcomings of the transformer-based object detection in local feature extraction, this article developed a DIAG-TR architecture. Three main components of DIAG-TR are specified as follows.

1) We designed a GLFI module upon a dual network structure in the encoder stage of the transformer architecture. The GLFI can mitigate the incompatibility of the global and local features, by feeding the local feature hierarchy with localization information to the self-attention branch, while feeding the global feature with context information to the CNN branch to realize a mutually enhancement (see Sections III-A and III-B).

2) After then, the coupled global-local features derived from the encoder stage are fed to the decoder stage, in which a content query and an anchor-box query are designed to probe the potential objects on the input feature map (see Section III-C).

3) Finally, bipartite matching method is adopted to automatically match the predicted object category as well as its anchor box with the ground truth (see Section III-D).

The main framework of the proposed method is presented in Fig. 1.

A. GLFI-Based Dual Network Structure in Encoder Stage

In order to embed the local feature hierarchy into the manifold of the global representation, and compensate for the incompatibility of the global and local features, the GLFI module based on dual network structure is designed, which consisting CNN branch and self-attention branch. Considering that local feature is a map structure with size of \( h \times w \times ch \) (where \( h, w, ch \) is the height, width, channel size of feature map), while the global feature is a sequential structure with size of \( d \times N \) (where \( d = S \times S \times ch \) is the channel size of each token vector, \( N = (h/S) \times (w/S) \) denotes the number of tokens, \( S \) denotes the patch size when generating tokens from the input image), the feature reconstruction mechanism [53] is adopted.
in GLFI to reconstruct the heterogenous features to accord with each other and mutually exchange the features between the CNN branch and self-attention branch, as shown in Fig. 2.

Taken the \(L\) layer GLFI for example, the input global and local feature is \(G_{L-1} \in R^{d \times N}\) and \(F_{L-1} \in R^{h \times w \times c}\), respectively. First, two convolution layers with \(1 \times 1\) and \(3 \times 3\) kernel (64 channel, and 1 stride, the following convolution layer is the same) are used to extract the intermediate local feature \(F'_{L-1}\), which is then reconstructed to a sequence structure by feature resampling process, and it is integrated to the self-attention branch by adding with \(G_{L-1}\)

\[
G_L = S\text{Block}
\]

\[
\times (\text{reshape}(W_{lf,1} \otimes F'_{L-1} + b_{lf,1}))
\]

(1)

where \(S\text{Block}\) denotes a self-attention unit. The feature resampling process consists of one \(1 \times 1\) convolution layer with weight \(W_{lf,1}\) and bias \(b_{lf,1}\) for cross-feature linear projection, the “down” operation \(\text{reshape}(\cdot)\) that reshapess the local feature from size \(d \times N\) to size \(h \times w \times c\), and the layer normalization \(\text{LayerNorm}(\cdot)\) layer that transform local feature to the statistical distribution of the global representation (see Fig. 2). \(\otimes\) is the convolution operation.

After deriving the embedded global feature \(G_L\), the feature resampling process is again adopted to inversely transform the \(G_L\) to the CNN feature structure, which is then integrated to the CNN branch by adding with the intermediate local feature \(F'_{L-1}\). The integrated features are finally passed to two convolution layers with \(1 \times 1\) and \(3 \times 3\) kernel (unifomrly represented by weight \(W_{lf}\) and bias \(b_{lf}\)), to obtain the embedded local feature \(F_L\). This process can be formulated as

\[
F_L = W_{lf} \left( F'_{L-1} + \text{BatchNorm} \right)
\]

\[
\times (W_{gf,1} \otimes \text{reshape} \left( G_L + b_{gf,1} \right)) + b_{lf}
\]

(2)

In the inverse direction, the feature resampling process consist of “up” operation \(\text{reshape}(\cdot)\) that reshapess the global feature from size \(d \times N\) to size \(h \times w \times c\), one convolution layer with weight \(W_{gf,1}\) and bias \(b_{gf,1}\) for cross-feature linear projection, and a batch normalization layer \(\text{BatchNorm}(\cdot)\) that transform the global feature to the distribution of local feature (see Fig. 2).

Finally, the embedded local feature \(F_L\) and the embedded global feature \(G_L\) are transmitted to the \(L + 1\) layer GLFI.

### B. Global Feature Representation With Self-Attention Branch

The specific structure of self-attention unit in GLFI is introduced in this section. Global feature representation can reflect the association or similarity between discretely and heterogeneously distributed geospatial object in a large receptive field perspective, which has latent spatial context relationship that is benefit to infer the category and location of the ground objects. The global feature extraction is implemented by a self-attention structure, as shown in Fig. 2.

The processing unit of the self-attention layer is a sequence of tokens, thus, the patch partition preprocessing is taken to reform the image to accord with the fashion of tokens. Given the input remote sensing image \(Y \in R^{h \times w \times c}\), patch partition is implemented to divide the image to a permutation of patches \(p_i \in R^{S \times S \times c}, i = 1, \ldots, N\) with size of \(S \times S\), and the total number of patches is \(N = (h \times w)/(S \times S)\). Afterwards, each patch is flattened to a token vector \(y_i \in R^{d \times 1}, i = 1, \ldots, N\), where \(d = S \times S \times c\). Each token is linearly projected to a patch embedding \(T = \{t_i, i = 1, \ldots, N\}\) by a learnable transformation matrix \(W_i \in R^{d \times d}\)

\[
t_i = W_i y_i, \quad i = 1, \ldots, N.
\]

Since the underlying implementation of self-attention is disordered permutation, the way of splitting image to patches may lose the original spatial location information of each token. To this end, we adopted an absolute spatial positional encoding to store the location information (see Fig. 2), and each element of the position encoding (PE) can be given as

\[
e_i(2j) = \left[ \sin \left( i/Temperature^{2j/d} \right) \right], j = 1, \ldots, d/2
\]

\[
e_i(2j + 1) = \left[ \cos \left( i/Temperature^{2j/d} \right) \right], j = 1, \ldots, d/2
\]

(4)

where \(e_i\) denotes the encoding of the \(i\)th token, \(e_i(2j)\) and \(e_i(2j + 1)\) denotes the odd and even encoding values in \(e_i\). The Temperature is empirically set to 20 [17].

The positional encoding \(e_i\) is elementwisely added with the corresponding patch embedding \(W_i y_i\), to integrate position information

\[
t_i = W_i y_i + e_i, \quad i = 1, \ldots, N.
\]

(5)

Based on the attention aggregation mechanism [54], the query \(Q = \{q_i, i = 1, \ldots, N\}\), key \(K = \{k_i, i = 1, \ldots, N\}\) and value \(V = \{v_i, i = 1, \ldots, N\}\) are first generated by linearly
transformation of patch embedding $t_i$ with learnable transformation matrix $W^q \in \mathbb{R}^{d \times d}$, $W^k \in \mathbb{R}^{d \times d}$, $W^v \in \mathbb{R}^{d \times d}$,

$$q_i = W^q t_i, k_i = W^k t_i, v_i = W^v t_i$$

$i = 1, \ldots, N$. (6)

Given a token $t_i$ to be retrieved, the attention aggregation mechanism enables it to search in global space to establish the relationship between $t_i$ and the residual token $t_j$ (where $j = 1, \ldots, N$), by matching the query of $t_i$ with the key of $t_j$

$$\hat{\alpha}_{i,j} = \text{SoftMax} \left( k_j^T q_i / \sqrt{d} \right)$$

$i = 1, \ldots, N, j = 1, \ldots, N$ (7)

where $\hat{\alpha}_{i,j}$ represent the spatial correlation (similarity) between $t_i$ and $t_j$, $k_j^T$ is the transpose of $k_j$, $1/\sqrt{d}$ is the normalization factor.

Taken the spatial correlation as cue, all the feature embeddings are weighted by multiplying the value of the $i$th token $v_i$ with the corresponding spatial correlation $\hat{\alpha}_{i,j}$, to guides the network to focus on the features that has distinct contributions to the spatial reasoning of the object distribution. The derived global feature representation for token $t_i$ can be given as

$$h_i = \sum_{j=1}^{N} v_i \hat{\alpha}_{i,j}, j = 1, \ldots, N.$$ (8)

Equation (8) can be given by matrix form

$$H = \text{Attention}(Q, K, V) = V \cdot \text{SoftMax}(K^T \cdot Q / \sqrt{d})$$ (9)

where $Q \in \mathbb{R}^{d \times N}$ represent the total queries of the tokens to be retrieved, $K \in \mathbb{R}^{d \times N}$, $V \in \mathbb{R}^{d \times N}$ represent the total keys and values of all the tokens, and $H \in \mathbb{R}^{d \times N}$ represent the extracted global feature representation. However, single stage self-attention has limitation for the reason that it only learns information in a single set of learnable transformation matrices. So, we adopt multihead self-attention to learn different representation by different learnable transformation matrices $W_h \in \mathbb{R}^{d \times d}, h = 1, \ldots, NH$, which can be given as

$$H_h = \text{Attention}(QW^q_h, KW^k_h, VW^v_h), h = 1 \ldots NH$$ (10)

where $NH$ is the total number of heads, $H_h$ is the attention matrix of the $h$th head.

And the final output $B \in \mathbb{R}^{d \times N}$ of the multihead self-attention unit is achieved by concatenating all heads

$$B = \text{Concat}(H_1, \ldots, H_T) W^O$$ (11)

where $W^O$ is the transformation matrix that aims to shrink the feature dimension of the concatenated head to $d$.

Generally, the input images are usually in the form of batch processing, and a layer normalization (LayerNorm) operation is implemented for one batch of the images, to normalize the images to Gaussian distribution with zero mean and variation 1

$$\text{LayerNorm}(B) = g_{ln} \times \frac{B - \mu(B)}{\sigma(B)} + b_{ln}$$ (12)

where $g_{ln}$ and $b_{ln}$ is the learnable scale and bias.
After then, the feed forward network (FFN) with multiple layer perceptron (MLP) is adopted to enhance the fitting capabilities

\[
MLP(B) = W_{ffn2} \odot \sigma(W_{ffn1} \odot B + b_{ffn1}) + b_{ffn2}
\] (13)

where \(W_{ffn1}, W_{ffn2}, b_{ffn1}, b_{ffn2}\) are weight and bias, \(\sigma\) is the ReLU activation function. Here, MLP is achieved by \(1 \times 1\) convolution layer. Besides, a residual structure is adopted for a robust learning, which can be given as

\[
O_1 = \text{LayerNorm}(T + B)
\]

\[
G_L = \text{LayerNorm}(O_1 + \text{MLP}(O_1)).
\] (14)

After multiple GFLI module, the output of the global representation branch (denoted as \(G\)) is taken as the output of the encoder part and is fed to the decoder part.

C. Decoder Stage

The role of the decoder part is to find out the desired object and its locations on the input feature \(G\) through an object query. The object query is a learnable sequential that deposit the information concerning the interest object, with its category and bounding box coordinates. Specifically, the cross-attention unit is used in the decoder layer to establish the association between decoder object query and the input feature \(G\), which is equivalent to probe the potential object and its location on the input features with similar morphology and region with the decoder object query.

Indeed, the original decoder object query adopted in DETR is commonly initialized to zero, without any inductive bias or prior, which makes the DETR extremely difficult to learn a proper object query, and results in a very slow convergence. To this end, we decouple the object query to content query (category information) and positional query (location information), in which a learnable anchor box is embedded in the positional query to provide a spatial prior [17]. The detail structure of one decoder layer is provided in Fig. 3, which usually consists of one self-attention unit, one cross-attention unit and FFN.

Given the \(i\)th anchor \((x_i, y_i, h_i, w_i)\) in the anchor box sequential \((N_i, \text{number of anchor box is predefined}, x_i, y_i)\) is the center coordinate, \(h_i, w_i\) is the height and width, the positional query of the \(i\)th anchor can be derived by anchor sine encoding \(\text{PE}()\) and MLP

\[
p_i = \text{MLP} \left( \text{Concat} (\text{PE}(x_i), \text{PE}(y_i), \text{PE}(w_i), \text{PE}(h_i)) \right)
\] (15)

where \(\text{PE}()\) operation is provided in (4), which can map a scalar value to a vector with size \(d/2\). \(\text{Concat}()\) operation is to concatenate the four vectors. \(\text{MLP}()\) consist of a linear layer and ReLU activation layer, which is used to reduce the feature dimension from \(2d\) to \(d\). Therefore, the dimension of positional query is \(P \in R^{d \times N_i}\).

In the multilhead self-attention module of the decoder part, the query \(Q_d \in R^{d \times N_i}\), key \(K_d \in R^{d \times N_i}\), and value \(V_d \in R^{d \times N_i}\) of the content query \(C \in R^{d \times N_i}\) are derived, as shown in (6). Different as before, since the content query itself has no positional information, we additionally integrate the positional query \(P\) to \(Q\) and \(K\), since the query and key are two main variables that mutually retrieval to derive the attention map. The formulation of \(Q_d, K_d, V_d\) can be given as

\[
Q_d = W^qC + P, K_d = W^kC + P, V_d = W^vC.
\] (16)

The attention aggregation operation is implemented as (9)–(11) to derive the embedded content query \(C_1 \in R^{d \times N_i}\).

In the cross-attention module, the content query and positional query are simultaneously used to search on the features \(G \in R^{d \times N}\). Specifically, the content query \(C\) is integrated into the positional embedding \(P\) by elementwise multiplication to provide target information, which is then concatenated with the transformed \(C_1\) to serve as the query \(Q_C\) for subsequent cross-attention. Therefore, the size of \(Q_C\) is expanded to \(2d \times N\). Besides, the content query \(C\) is firstly rescaled by MLP to transform it to the distribution of the positional embeddings \(P\). The formulation of \(Q_C\) can be expressed as

\[
Q_C = \text{Concat} (W^qC_1, P \cdot \text{MLP}(C))
\] (17)

where \(\cdot\) denotes elementwise multiplication.

Since the purpose of decoder is to explicitly explore the similarity between query (from decoder) and the feature (from encoder), the key \(K_C\) and value \(V_C\) of the cross-attention module is derived from the feature \(G\) and the spatial positional encodings \(SP \in R^{d \times N}\) used in the encoder part. Specifically, the key \(K_C\) is derived by concatenating the transformed \(G\) with spatial positional encodings \(SP\), and the size of \(K_C\) is expanded to \(2d \times N\), which can be expressed as

\[
K_C = \text{Concat} (W^kG, SP).
\] (18)
The value \( V_C \) is derived by the transformed \( G \)

\[
V_C = W^v G. \tag{19}
\]

In this way, the attention aggregation operation is also implemented to derive the embedded content query \( C_2 \).

Finally, a FFN with MLP and a residual connection is adopted to transform \( C_2 \) to a decoder embedding to feed it to the next decoder layer. In order to modulate the anchor box, the FFN additionally transformed \( C_2 \) to the gradient \((\Delta x, \Delta y, \Delta h, \Delta w)\) for updating by linear MLP, and the updated anchor box are also fed to the next decoder layer. The final decoder layer outputs the content embeddings \( CE \in R^{d \times N_i} \), and anchor box \( AB \in R^{d \times 4} \), where \( CE \) is fed to a MLP classifier to derive the category probability \( CP \in R^{NC \times N_i} \) (\( NC \) denotes the number of categories).

\section{B. Bipartite Matching}

Instead of traditional loss function calculation that use pair of prediction (category and bounding box) and ground truth defined by the region-proposal/anchors for training, the transformer-based method regarded it as a disorder set prediction and matching process, which automatically match the prediction with its ground truth based bipartite matching.

Specifically, \( M \) ground truth object regions are in the input image, and \( N_i \) objects are predicted from the decoder part (generally \( M \leq N_i \)), thus, there are totally \( M \times N_i \) possible matching cases, and bipartite matching is adopted to automatically select the optimal matching for loss function calculation.

Supposing the \( i \)-th ground truth object \( o_i \) is matching with the \( f(i) \) predicted object \( \hat{o}_{f(i)} \) (\( f(\cdot) \) is the matching mapping), our purpose is to minimize the matching loss \( L_{\text{match}} \)

\[
\hat{f} = \arg \min_{f \in \mathcal{S}} \sum_{i=1}^{M} L_{\text{match}} (o_i, \hat{o}_{f(i)}) \tag{20}
\]

where \( \hat{f} \) is the estimated optimal matching. The matching process are two-fold, i.e., category probability and anchor box. Specifically, Hungarian matching function \( L_H \) is adopted

\[
L_H (o, \hat{o}) = \sum_{i=1}^{M} \left[ - (cp_i) \log c\hat{p}_{f(i)} + 1_{\{cp_i \neq 0\}} L_{\text{box}} (ab_i, \hat{a}_{f(i)}) \right] \tag{21}
\]

where \( cp_i \in R^{NC \times 1} \) and \( ab_i \in R^{1 \times 4} \) are the probability vector and anchor box of the \( i \)-th ground truth object, while \( \hat{c}\hat{p}_{f(i)} \) and \( \hat{a}_{f(i)} \) are possible corresponding predicted object. The first term is a cross entropy, and the second term can be expressed as

\[
L_{\text{box}} (ab_i, \hat{a}_{f(i)}) = \lambda_1 L_{\text{iou}} (ab_i, \hat{a}_{f(i)}) + \lambda_2 \| ab_i - \hat{a}_{f(i)} \|_2 \tag{22}
\]

where \( L_{\text{iou}} (\cdot) \) is the intersection over union loss function between two regions, \( \lambda_1 \) and \( \lambda_2 \) are predefined parameters.

\section{IV. Experiments and Analysis}

\subsection{A. Dataset}

DIOR dataset [7], which is the largest optical remote sensing image dataset for object detection community in both the category and instance aspect is adopted for validation experiment. DIOR dataset contains 23462 images, and 192472 object instances over 20 common object categories, including airplane (AL), airport (AT), baseball field (BF), basketball court (BC), bridge (B), chimney (C), dam (D), expressway service area (ESA), expressway toll station (ETS), golf course (GC), ground track field (GT), harbor (HB), overpass (O), ship (S), stadium (SD), storage tank (ST), tennis court (TC), train station (TS), vehicle (V), and wind mill (WM). Each category contains roughly 1200 images. The image size is 800 \times 800 pixels and the spatial resolution is 0.5–30 m. In order to compensate for the diverse imaging conditions in earth observation, the images are obtained with different weathers, seasons, quality, and has high interclass similarity and intraclass diversity.

\subsection{B. Implementation Details}

During the training process, AdamW Optimizer is adopted with a basic learning rate of \( 10^{-5} \) for backbone, \( 10^{-4} \) for other parts, a momentum coefficient of 0.9, and a weight decay of \( 10^{-4} \). In the initial training stage, training process is implemented with the initial settings. After training 40 epochs, learning rate would drop tenfold for better optimization. Encoder and decoder depth is set to 12 and 6, respectively, to maximize the global and local representation in DIAG-TR. We achieved the highest accuracy on DIOR dataset when patch size \( S \) is set to 16 in the self-attention branch in GLFI, the embedding dimension \( d \) is set to 256 in both encoder and decoder, and the number of object queries \( N_i \) is 300. \( \lambda_1 \) and \( \lambda_2 \) in \( L_{\text{box}} (ab_i, \hat{a}_{f(i)}) \) is set to 2, 5, respectively. The experiments are performed on a server with RTX 3090 graphics processing unit accelerators of total 72 GB memory. The environment adopted Python 3.8 and pytorch 1.8.0.

For making a comparison, ten state-of-the-art methods are selected as benchmark, including region-proposal-based methods like R-CNN [8], RICNN [26], Faster RCNN [23], Faster RCNN with FPN [14], Mask R-CNN with FPN [24], PANet [33], regression-based methods like SSD [40], YOLOv3 [39], RetinaNet [42], and transformer-based method like DETR [15].

To evaluate the proposed DIAG-TR, general object detection criterion mean average precision (mAP) is used for evaluation metrics. Precision can be formulated as following:

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{23}
\]

where TP denotes the number of correctly detected objects, namely true positives (TPs). Conversely, FP denotes the number of incorrectly detected objects, namely false positives (FPs). In order to determine whether the detected target is correct, intersect over union (IoU) is widely used in object detection task. Different IoU settings will significantly influence the value of mAP. Therefore, we can define \( mAP_i \) as mean average precision when IoU between category-correct prediction box and ground truth box is over \( i \). Similarly, \( mAP_{ij} \) denotes as the average \( mAP_i \) where \( i \) is from \( i \) to \( j \) with interval of 5.
C. Comparison Results With State-of-the-Art Method

Fig. 4 shows the visual inspection results of the proposed DIAG-TR on DIOR dataset. It is notable that our proposed DIAG-TR is able to detect large-scale, median-scale, and small-scale simultaneously. Especially on the objects that perform strong spatial dependency with the background, i.e., vehicles in the express service area or vehicles nearby the express toll station. More specifically, cars which are heavily blocked by the buildings can also be detected by DIAG-TR, which means our inference results have exhibited a satisfactory detection capability on detecting small-scale objects in a complex background. Meanwhile, by coupling hierarchy inductive bias from CNN with transformer encoder, same object class on different scales, such as huge warships and small yachts are both well visualized in the inference results.

In order to further confirm the ability of our model in global context modeling and hierarchical inductive bias, Fig. 5 shows an intuitive comparison of detection results. In Fig. 5, our proposed DIAG-TR can not only successfully detect storage tanks with white top, but also storage tanks with black top. However, the
result of DETR can only inference white storage tanks. With better ability to obtain a larger receptive field by coupling transformer and CNN in GLFI, all small targets vehicles in express service area with ground color and an overpass can be correctly identified by DIAG-TR while DETR omits some vehicles and wrongly detects the overpass as bridge.

For the quantitative assessment, we calculate $mAP_{50}$ and $AP_{50}$ for each class of SOTA model and our proposed DIAG-TR. In Table I, we can see transformer-based model (DETR and DIAG-TR) perform a noticeable advantage compared to the traditional CNN-based model. In BC, C, ETS, we achieved the best results in $AP_{50}$ and also achieved near SOTA performance in AT, GC, SD, and TS. Therefore, our model generally has excellent ability to detect large-scale and median-scale objects. However, DETR has an obvious decline in ETS, TC, and SD due to simple CNN-based ResNet-50 backbone and object query without position information. Besides, from the comparison between DETR and DIAG-TR, it is obvious that the proposed GLFI module contributes to 3.4% $mAP_{50}$ improvement. When comparing with the other state-of-the-art methods, it is found that DIAG-TR outperforms region-proposal-based Faster-RCNN-FPN by 8.9%, and outperforms regression-based RetinaNet by 6.3%.

For evaluating the computation complexity of our model, we measure the values of floating-point operations (FLOPs) and Parameters (Params) of DIAG-TR and baseline methods in Table II. Although we achieved highest mAP at 72.0 with the inputs size of 800 × 800 in DIOR dataset, the computation complexity is still expected to be improved in the future due to the transformer design.

D. Ablation Study

In order to validate the effectiveness of the proposed DIAG-TR, we perform 3 sets of ablation experiments on DIOR dataset, to explore each key components and parameter setting of DIAG-TR. The details of the experiments are as following.

1) Analysis of Different Depths of DIAG-TR Encoder: In Table III, with encoder going deeper by increasing the number of GLFI modules, the accuracy results gradually increase 1.4 $mAP_{50}$, 1.1 $mAP_{75}$, 0.8 $mAP_{50:95}$ and 0.8 $mAP_{50}$, 1.1 $mAP_{50:95}$ from value of 3 to 6 and from value of 6 to 12, respectively. That is because GFLI module possess powerful global and local representation modeling ability. At the same time, hierarchical structure and residual connection make the model possible to build deeper and improve the accuracy through multiscale feature learning.

2) Analysis of DIAG-TR Decoder: We replace the DETR backbone and encoder with a single DIAG-TR encoder to explore the advantages of the DIAG-TR decoder. We find that

---

**TABLE I**

| Method       | R-CNN | RCNN | Faster RCNN | SSD | YOLOv3 | Faster RCNN with RPN | Mask RCNN with FPN | RetinaNet | PANet | DETR | DIAG-TR |
|--------------|-------|------|-------------|-----|--------|---------------------|-------------------|-----------|-------|------|---------|
| Backbone     | VGG16 | VGG16| VGG16       | VGG16| YOLOv3 | VGG16               | VGG16             | YOLOv3    |       |      |         |
| mAP          | 37.7  | 44.2 | 54.1        | 58.6| 57.1   | 63.1                | 63.5             | 65.7      | 63.8  | 68.6 | 72.0    |

**TABLE II**

| Method       | Paras(M) | FLOP(G) |
|--------------|-----------|---------|
| SSD          | 185.1     | 82      |
| YOLOv3       | 240.1     | 71      |
| Faster RCNN with FPN | 42.72 | 180 |
| Mask RCNN with FPN | 45.91 | 199 |
| RetinaNet    | 38.64     | 205     |
| PANet        | 25.13     | 449     |
| DETR         | 54.73     | 152     |
| DIAG-TR      | 61.42     | 190     |

**TABLE III**

| Encoder depth | $mAP_{50}$ | $mAP_{75}$ | $mAP_{50:95}$ |
|---------------|------------|------------|--------------|
| none          | 68.3       | 36.6       | 38.2         |
| sine          | 72.0       | 42.1       | 43.0         |

**TABLE IV**

| Positional Encoding | $mAP_{50}$ | $mAP_{75}$ | $mAP_{50:95}$ |
|---------------------|------------|------------|--------------|
| none                | 60.3       | 36.6       | 38.2         |
| sine                | 72.0       | 42.1       | 43.0         |
the decoder advantage of DIAG-TR is mainly manifested in the model training process. In Fig. 6, we can see that DIAG-TR converges on the training set and validation set at around 180 epochs, but DETR with GLFI encoder needs around 450 epochs to achieve the lowest loss value, which shortens the convergence time by 2.5-fold. The reason is our decoder assigns a position prior to the decoder object query, so that the decoder can quickly focus on a certain area, while the DETR query is directly initialized to 0, which is spatially mapped to a fixed certain area. Therefore, the DETR query is more difficult to learn the position information.

3) Analysis of Spatial Positional Encoding: Self-attention computing cannot reflect the location relationship between two tokens, so it is necessary to add positional encoding to enhance location information in both encoder and decoder. Especially in the task of object detection, which has accurate demand for target location, absolute positional encoding is very important. Therefore, the results in Table IV show that the DIAG-TR with positional encoding has a great improvement of +3.7 mAP$_{50}$, +5.5 AP$_{75}$, and +4.8 mAP$_{50:95}$.

V. CONCLUSION

This article explores the feasibility of the novel transformer-based object detection architecture on reasoning global spatial relation of the sparsely and heterogeneously distributed geospatial objects in remote sensing image, and focuses on mitigating its weakness of modeling local feature hierarchy and slow convergence problem. In the proposed DIAG-TR, a GLFI module is designed to uniform the paradigm of global and local feature, and hierarchically embed the local features into global representations. Besides, a learnable anchor box is incorporated into the positional query in the decoder part to provide a spatial prior to accelerate convergence. From experiment results validated on the optical remote sensing image dataset DIOR, we find that the DIAG-TR has better spatial reasoning ability on the objects that exhibit strong spatial dependency with the background, such as vehicles in the express service area or nearby the express toll station. Meanwhile, DIAG-TR also shows a satisfied detection capability on small-scale objects in complex background, like cars around the buildings. However, in remote sensing community, object detection is facing the challenges of cloud contamination, which is also a research topic of our study in future [55], [56]. We hope the findings explored in this article can provide insight into a better understanding of the transformer paradigm on object detection method, and further facilitate object detection in earth observation community.

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] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, “Building detection in very high resolution multispectral data with deep learning features,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1873–1876.
[3] Z. Chen, T. Zhang, and C. Ouyang, “End-to-end airplane detection using transfer learning in remote sensing images,” *Remote Sens.*, vol. 10, 2018, Art. no. 139.
[4] Z. Zou and Z. Shi, “Ship detection in spaceborne optical image with SVD networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5832–5845, Oct. 2016.
[5] Z. Chen et al., “Vehicle detection in high-resolution aerial images via sparse representation and superpixels,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 103–116, Jan. 2016.
[6] G. Cheng and J. Han, “A survey on object detection in optical remote sensing images,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 117, pp. 11–28, 2016.
[7] 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. 139, pp. 296–307, 2019.
[8] 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.
[9] Z. Deng et al., “Multi-scale object detection in remote sensing imagery with convolutional neural networks,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 3–22, 2018.
[10] 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.
[11] X. Han, Y. Zhong, and L. Zhang, “An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery,” *Remote Sens.*, vol. 9, no. 7, 2017, Art. no. 666.
[12] 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.
[13] 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.
[14] T. Y. Lin et al., “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
[15] N. Carion et al., “End-to-end object detection with transformers,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[16] X. Zha et al., “Deformable DETR: Deformable transformers for end-to-end object detection,” 2020, arXiv:2010.04159.
[17] S. Liu et al., “DAB-DETR: Dynamic anchor boxes are better queries for DETR,” 2022, arXiv:2201.12329.
[18] 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, Aug. 2022, Art. no. 4409512.
[19] 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.
[20] 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.
[21] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. Conf. Adv. Neural Inform. Process. Syst., 2016, pp. 379–387.

[22] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1440–1448.

[23] 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.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2961–2969.

[25] K. Li, G. Cheng, S. Bu, and X. You, “Rotation-insensitive and context-augmented object detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[26] 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.

[27] G. Cheng, J. Han, P. Zhou, and D. Xu, “Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 265–278, Jan. 2019.

[28] Y. Zhong, H. Han, and L. Zhang, “Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery,” ISPRS J. Photogrammetry Remote Sens., vol. 138, pp. 281–294, 2018.

[29] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, “Accurate object localization in remote sensing images based on convolutional neural networks,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 5, pp. 2486–2498, May 2017.

[30] Z. Deng et al., “Multi-scale object detection in remote sensing imagery with convolutional neural networks,” ISPRS J. Photogrammetry Remote Sens., vol. 145, pp. 3–22, 2018.

[31] W. Guo, W. Yang, H. Zhang, and G. Hua, “Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network,” Remote Sens., vol. 10, 2018, Art. no. 131.

[32] X. Zhang et al., “Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network,” Remote Sens., vol. 11, 2019, Art. no. 755.

[33] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8759–8768.

[34] Z. Xu, X. Xu, L. Wang, R. Yang, and F. Pu, “Deformable convnet with aspect ratio constrained NMS for object detection in remote sensing imagery,” Remote Sens., vol. 9, 2017, Art. no. 1312.

[35] J. Dai et al., “Deformable convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 764–773.

[36] T. Tang, S. Zhou, Z. Deng, L. Lei, and H. Zou, “Arbitrary-Oriented vehicle detection in aerial imagery with single convolutional neural networks,” Remote Sens., vol. 9, 2017, Art. no. 1170.

[37] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788.

[38] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6517–6525.

[39] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.

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

[41] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DSSD: Deconvolutional single shot detector,” 2017, arXiv:1701.06599.

[42] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, Feb. 2020.

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

[44] W. Liu, L. Ma, and H. Chen, “Arbitrary-Oriented ship detection framework in optical remote-sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 6, pp. 937–941, Jun. 2018.

[45] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, “Automatic ship detection based on retinanet using multi-resolution gaofen-3 imagery,” Remote Sens., vol. 11, no. 5, 2019, Art. no. 531.

[46] U. Balganci, M. Soydas, and E. Sertel, “Comparative research on deep learning approaches for airplane detection from very high-resolution satellite images,” Remote Sens., vol. 12, 2020, Art. no. 458.

[47] J. Beal et al., “Toward transformer-based object detection,” 2020, arXiv:2012.09958.

[48] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, “Rethinking transformer-based set prediction for object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3611–3620.

[49] X. Xu, Z. Feng, C. Cao, M. Li, and J. Wu, “An improved swin transformer-based model for remote sensing object detection and instance segmentation,” Remote Sens., vol. 13, 2021, Art. no. 4779.

[50] T. Ma et al., “ Oriented object detection with transformer,” 2021, arXiv:2106.03146.

[51] F. Li et al., “DN-DETR: Accelerate DETR training by introducing query denoising,” 2022, arXiv:2203.01305.

[52] H. Zhang et al., “DINO: DETR with improved denoising anchor boxes for End-to-End object detection,” 2022, arXiv:2203.03605.

[53] Z. Peng et al., “Conformer: Local features coupling global representations for visual recognition,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 367–376.

[54] A. Vaswani et al., “Attention is all you need,” Adv. Neural Inf. Process. Syst., vol. 30, 2017, Art. no. 48913.

[55] X. Li, R. Feng, X. Guan, H. Shen, and L. Zhang, “Remote sensing image mosaicking: Achievements and challenges,” IEEE Geosci. Remote Sens. Mag., vol. 7, no. 4, pp. 8–22, Dec. 2019.

[56] Z. Li, L. Wang, Q. Cheng, P. Wu, W. Gan, and L. Fang, “Cloud removal in remote sensing images using nonnegative matrix factorization and error correction,” ISPRS J. Photogrammetry Remote Sens., vol. 148, pp. 103–113, 2019.

Jingqian Xue is currently working toward the B.S. degree in geographic information science with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. His research interests include remote sensing image processing, geographically explainable deep learning, and subpixel mapping.

Da He (Member, IEEE) received the B.S. degree in remote sensing science and technology and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2015 and 2020, respectively. He is currently a Postdoctor with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. His current research interests include multi- and hyperspectral remote sensing image classification, deep learning, data fusion, and change detection.

Mengwei Liu received the B.S. degree in computer science and technology from Zhengzhou University, Zhengzhou, China, in 2005, and the M.S. and Ph.D. degrees in hydrology and water resources and geographical information system information from Sun Yat-sen University, Guangzhou, China, in 2007 and 2010, respectively. He is currently the Guangxi Academy of Science Spatial Information Team Deputy Researcher, Guangzhou Survey Geographic Information Association expert. His current research interests include spatial data mining, multiobjective space optimization, photogrammetry and remote sensing, geographic information system theoretical research, and technical work.

Qian Shi (Senior Member, IEEE) received the B.S. degree in sciences and techniques of remote sensing from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, in 2015. He is currently an Associate Professor with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. Her research interests include remote sensing image classification, including deep learning, active learning, and transfer learning.