Efficient Inductive Vision Transformer for Oriented Object Detection in Remote Sensing Imagery

Cong Zhang, *Graduate Student Member, IEEE*, Jingran Su, Yakun Ju, *Member, IEEE*, Kin-Man Lam, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Object detection is a fundamental task in remote sensing image analysis and scene understanding. Previous remote sensing object detectors are typically based on convolutional neural networks (CNNs), whose performance is significantly limited by the intrinsic locality of convolution operations. The emergence of vision Transformers brings potential solutions to this problem, which has the capability to be a solid alternative to CNNs. However, three crucial obstacles hinder the application and performance of Transformers in the task of remote sensing object detection, that is: 1) high computational complexity, especially for high-resolution remote sensing images; 2) training and sample inefficiency caused by lack of inductive bias; and 3) difficulty in learning arbitrary orientation knowledge of geospatial objects. To address these issues, in this article, a novel efficient inductive vision Transformer framework is proposed for oriented object detection in remote sensing imagery. This framework follows the hierarchical feature pyramid structure and makes threefold contributions as follows: 1) spatial redundancy in remote sensing images is fully explored and an adaptive multigrained routing mechanism is proposed to facilitate token sparsity in Transformers, which can dramatically reduce the computational cost without comprising the accuracy. 2) A compact dual-path encoding architecture, where both global long-range dependencies and local semantic relations are jointly and complementarily captured, is proposed to enhance inductive bias in Transformers. 3) An angle tokenization technique is proposed to promote the encoding, embedding, and learning of direction knowledge for oriented objects in remote sensing scenarios. In this work, the above-mentioned three contributions are instantiated in an advanced Transformer-based object detector, namely, EIA-pyramid vision Transformer (PVT). Comprehensive experiments on two publicly available datasets have demonstrated its effectiveness and superiority for oriented object detection in remote sensing images.**

***Index Terms*— Adaptive tokens, inductive biases, object detection, remote sensing imagery, vision Transformers.**

I. Introduction

OBJECT detection is one of the critical tasks in the field of remote sensing image analysis and has been extensively utilized in various real-world applications, such as smart cities, precision agriculture, traffic monitoring, and urban planning [1], [2], [3], [4]. The goal of object detection is to accurately locate and categorize specific objects in a given image. In the past decade, deep learning has played a vital role in the progress toward robust and precise object detection across diverse scenarios [5], [6], [7], [8]. In particular, with the exceptional representation ability, modern generic object detectors [9], [10], [11], [12], [13], [14], based on convolutional neural networks (CNNs), have achieved dominant detection performance in natural scenes, such as the PASCAL VOC [15] dataset and the MS COCO dataset [16]. Inspired by this achievement, numerous variants of the CNN-based detector have also been adapted specifically for remote sensing scenarios, presenting impressive performance gains [17], [18], [19]. However, more recently, the dominance of CNNs has been seriously challenged, due to the limitations of intrinsic locality brought by convolution operations. The self-attention mechanism [20] has been demonstrated to effectively tackle this issue by capturing nonlocal long-range dependencies, which further motivates the development of standard vision Transformers [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33].

As a state-of-the-art technique with more powerful discriminative ability, the vision Transformer has promising potential to be a solid alternative to CNNs for various high-level tasks in natural scenes, including object detection [34], [35], [36], [37], semantic segmentation [38], [39], [40], and visual tracking [41], [42]. In the context of Earth observation, several efforts have been made to adapt generic vision Transformers to specific remote sensing applications, such as change detection [43], road detection [44], and building extraction [45], all of which have reported improved performance compared with traditional CNN-based frameworks. Intuitively, the self-attention mechanism in Transformers is crucial for most tasks on remote sensing images, including satellite and aerial images. On the one hand, different from the natural scenes in a head-up view, optical remote sensing images usually follow a bird's-eye perspective, naturally with richer object content and semantic information, consequently, depending more on global long-range dependencies. On the other hand, the scale of geospatial objects varies considerably for different ground-sampling distances from aircraft, and thus modeling flexible scale-invariant relations in vision Transformers is preferred. The above-mentioned inherent properties rationalize the priority of Transformers over pure CNNs for remote

Manuscript received 15 March 2023; revised 13 May 2023; accepted 21 June 2023. Date of publication 5 July 2023; date of current version 7 August 2023. *(Corresponding author: Cong Zhang.)*

Cong Zhang, Yakun Ju, and Kin-Man Lam are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: cong-clarence.zhang@connect.polyu.hk).

Jingran Su is with the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong.

Qi Wang is with the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi'an 710072, China.

Digital Object Identifier 10.1109/TGRS.2023.3292418

sensing tasks, including remote sensing object detection. For example, recently, Zheng et al. [46] proposed a remote sensing detector, namely, ADT-Det, where a feature pyramid Transformer was designated to promote nonlocal semantic representation at different feature levels.

However, the superiority of vision Transformers is still to be explored for the task of remote sensing object detection, because the following three fundamental obstacles prevent existing methods from efficiently leveraging the Transformer architectures to defeat pure CNN-based object detectors.

1) *High Computational Costs:* Captured by sophisticated satellites or aerial cameras, remote sensing images are typically high-resolution images with a high amount of redundant and complicated details, as well as structural information [48]. Considering the pairwise relationships inside vision Transformers, the computational and memory costs of exhaustively establishing high-resolution feature representations are enormous, or even unacceptable, for most remote sensing object detectors in practice.
2) *Lack of Inductive Bias:* Different from convolutions, vision Transformers naturally and essentially deteriorate inductive biases in modeling locality and scale invariance of objects. Alternatively, in Transformer-based frameworks, intrinsic inductive biases have to be learned compulsively and implicitly from very large-scale training data and with long training schedules [30], [33], [49]. This sample-inefficient regime adversely degrades object detection performance, owing to scarce fully annotated remote sensing data.
3) *Arbitrary Orientation:* Objects in natural scene images are usually horizontal, while objects in remote sensing images have arbitrary orientation and are densely arranged [50], [51], [52], [53], [54], [55]. Due to cluttered background interference and imprecise discrimination, it is challenging for horizontal feature tokens in previous Transformer-based detectors to learn orientation or angle knowledge of geospatial objects.

Natural scenes and remote sensing scenes embody considerably distinct image content, where objects of interest to be detected vary in appearance and texture, despite the same category, such as the airplanes (APLs) depicted in Fig. 1(a) and (c). More importantly, Fig. 1(b) and (d) shows dramatic differences between the statistical distribution of the spatial occupation of objects in natural scene images and remote sensing images. It can be observed that in most natural scene images, objects tend to occupy over half of the entire image, while more than 70% of remote sensing images contain geospatial objects only occupying less than 10% of the space. This implies that considerable background redundancy commonly exists in remote-sensing images, which can be suppressed to reduce computational complexity. This preliminary observation motivates us to further rigorously and quantitatively explore spatial redundancy in remote sensing imagery to overcome the aforementioned obstacles, especially the high computational inefficiency of vision Transformers. Concretely, Fig. 2 shows the empirical analyses conducted to investigate the similarity of Transformer tokens in local regions, based on the Pearson correlation coefficient metric. Taking the famous Transformer-based detector pyramid vision Transformer (PVT) [23] as an example, we first train it on two different scenarios, i.e., the MS COCO dataset (natural scenes) and the DOTA dataset (remote sensing scenes), respectively, and then calculate the correlation coefficients among adjacent tokens inside each local 2-D feature region or patch. As shown in Fig. 2(b), neighboring Transformer tokens (or queries) learned from remote sensing images generally share substantial similarities, which indicates that most queries are eminently redundant in a local patch. Moreover, as shown in Fig. 2(c), the spatial redundancy follows distinctive patterns for different encoder layers. Therefore, to reduce computational load, each encoder layer should be considered separately, instead of treating them equally.

Based on the above-mentioned observations and empirical analysis, to address the three critical issues, in this article, we propose a novel **E**fficient **I**nductive vision Transformer framework with **A**ngle tokenization for oriented object detection in remote sensing images, namely, **EIA-Transformer**. In theory, it can be adapted to most Transformer-based remote sensing detectors, whereas for generality and convenience, we exploit the well-known PVT [23] as the baseline model and form EIA-PVT in this work. Fig. 3 depicts the overall structure of EIA-PVT, which is composed of four cascaded stages with multiscale context, each of which consists of three dedicated components, namely, pyramid patch embedding (PPE), efficient inductive encoders (EIEs), and an angle tokenization module (ATM). Specifically, the proposed PPE progressively introduces intrinsic inductive bias of scale invariance to the embeddings (*against the second obstacle*). EIE includes an adaptive multigrained router (AMGR) to prompt token sparsity and reduce the computational complexity (*against the first obstacle*), followed by an inductive dual-path modulator (IDPM) to jointly incorporate global dependencies and the local inductive bias (*against the second obstacle*). ATM aims to explicitly discriminate the orientations of geospatial objects and guides the angle learning of feature tokens (*against the third obstacle*). Comprehensive experiments were conducted on different publicly available datasets for oriented object detection in remote sensing images, i.e., DOTA [47] and DIOR-R [50]. Experimental results have validated the effectiveness and superiority of our method. The major contributions of this article can be summarized as follows.

1) Based on the exploration of spatial redundancy in high-resolution remote sensing images, an adaptive multigrained routing mechanism is proposed to dynamically determine the sparsification of Transformer tokens, where uninformative tokens are reorganized in coarse granularity, thereby dramatically improving detection efficiency.
2) A compact parallel encoding architecture is proposed by efficiently integrating convolutions with self-attention operations, which has the high capability of simultaneously modeling long-range semantic relations and inductive bias of locality and scale invariance. Moreover, inductive biases are enhanced and maintained

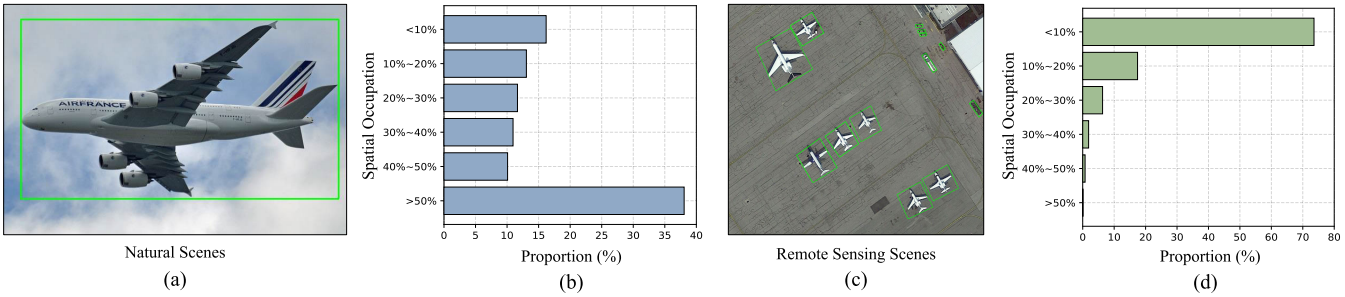


Fig. 1. Comparison of natural scenes and remote sensing scenes. (a) Natural scene image with detection annotations from the MS COCO dataset [16]. (b) Statistical distribution of spatial occupation of natural objects in the COCO dataset. The vertical axis represents the proportion of the space occupied by objects in an image, while the horizontal axis is the proportion of the number of corresponding images to the total number of images in the dataset. (c) Remote sensing image with detection annotations from the DOTA dataset [47]. (d) Statistical distribution of spatial occupation of remote sensing objects in the DOTA dataset. The axes have the same definitions as in (b).

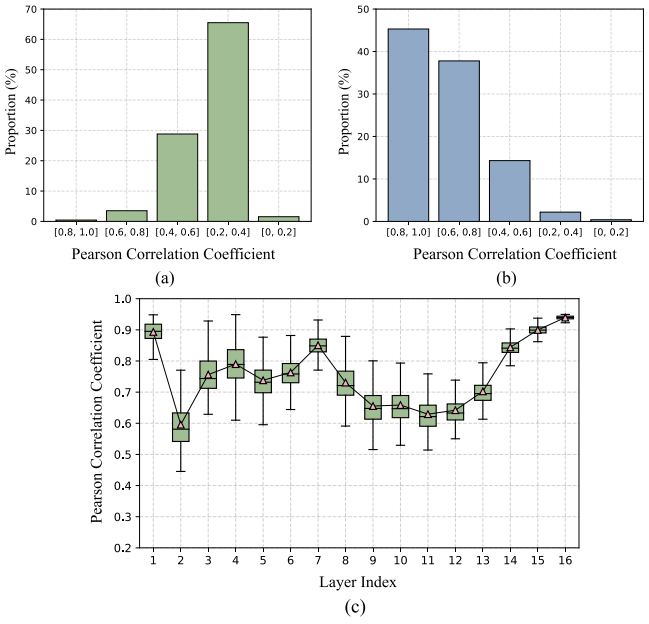


Fig. 2. Spatial redundancy statistics of the vanilla Transformer encoders in PVT-small [23]. (a) Proportion histogram of Transformer tokens against different ranges of Pearson correlation coefficients in the well-trained PVT on the MS COCO dataset. (b) Proportion histogram of Transformer tokens against different ranges of Pearson correlation coefficients in the well-trained PVT on the DOTA dataset. (c) Correlation coefficient statistics along different encoder layers inside the well-trained PVT on the DOTA dataset. Please note that a higher correlation typically implies more spatial redundancy.

throughout the Transformer encoder from different perspectives, to alleviate the issues of scale variations and sample inefficiency in remote sensing object detection.

3) A vision Transformer-friendly angle encoding, embedding, and learning paradigm is proposed for discriminative feature representations of remote sensing objects with arbitrary orientations, which delves into explicit and accurate supervision information for Transformer encoders at different feature levels.

The rest of this article is organized as follows. First, the related work is briefly reviewed in Section II. Then, Section III presents the proposed method in detail, and Section IV is focused on extensive experiments and result analyses. Finally, Section V draws the conclusion.

II. Related Work

Thanks to its promising applications in the real world, object detection has been extensively studied with impressive progress, in terms of both accuracy and efficiency. There have been frequent interactions between generic object detection and remote sensing object detection in the past decade, which will be briefly reviewed. Moreover, an emerging fundamental technology, highly related to this work, i.e., the vision Transformer, is also presented in this section.

A. Remote Sensing Object Detection

Generally, remote sensing object detection serves as a basic and primary task of various high-level applications, such as urban planning and resource exploration. Previous remote sensing object detection algorithms are based on traditional handcrafted feature extraction, such as saliency information [56], scale-invariant features [57], [58], and shape-texture features [59]. Limited by the unrobustness of handcrafted features, most conventional remote sensing detectors focus on the detection of a specific single category of geospatial objects, such as ship detection [60] and vehicle detection [61]. Recently, this deficiency has been significantly alleviated, due to the development of deep learning and CNNs, which have been widely utilized for generic object detection in natural scenes. As the most representative CNN-based detection framework, faster R-CNN [11] is based on an effective *two-stage* detection pipeline, where region proposals are first generated based on deep CNN features, followed by category prediction and bounding box regression for the region proposals. To improve model efficiency, *one-stage* detectors, such as YOLO [12] and RetinaNet [14], remove the region proposal stage and treat object detection as a straightforward regression problem, which can produce competitive performance and higher inference speed in natural images.

The success of CNN-based generic object detection has recently inspired an increasing number of CNN-based remote sensing object detectors [5], [62], [63], [64], [65], [66], [67]. For example, in [63], based on RetinaNet, an adaptive balanced network was devised to improve the detection accuracy of multiscale remote sensing objects, by introducing the enhanced channel attention mechanism and an adaptive feature pyramid network. However, unlike natural scene objects, geospatial

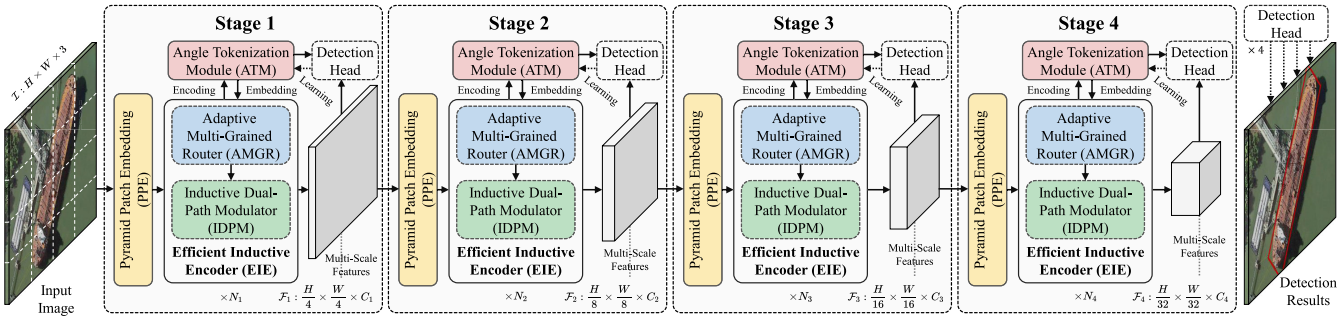


Fig. 3. Overview of the proposed EIA-PVT, which comprises four successive stages for multiscale representation generation. The different stages typically involve different numbers of channels for multiscale representations, and in practice, $\{C_1, C_2, C_3, C_4\} = \{64, 128, 320, 512\}$. Each stage mainly involves three key components, i.e., PPE, EIE, and ATM, while EIE is composed of AMGR and IDPM.

objects of interest in remote sensing images are usually densely distributed with arbitrary orientations. To address this issue, numerous oriented object detectors have been developed specifically for remote sensing scenarios [53], [54], [55], [68], [69], [70]. For instance, in [68], a two-stage dual-aligned oriented detector was designed for remote sensing object detection, which leveraged both oriented proposals and rotated RoI alignment modules to alleviate the spatial misalignment between horizontal proposals and oriented geospatial objects. Additionally, Yang et al. [69] presented a progressive regression approach to refine the one-stage detector for oriented objects at the feature level. Huang et al. [70] adopted 2-D oriented Gaussian heatmaps and developed an anchor-free object-adaptation label assignment strategy with an oriented-bounding-box representation component, to boost detection performance in remote sensing images. Although the remote sensing community has witnessed a remarkable growth in CNN-based object detectors, the local receptive field of convolution operations inherently limits their capacity to model long-range dependencies, especially in high-resolution optical aerial or satellite imagery. Accordingly, the community expects to obtain more advanced and powerful detection frameworks through the proper and efficient utilization of vision Transformers.

B. Vision Transformers

The concept of Transformers was first proposed in [20] to form a self-attention-driven machine translation model. After that, it rapidly replaced traditional recurrent neural networks in the field of natural language processing (NLP), due to its strong ability to handle long-distance encodings [71]. Furthermore, inspired by the success of Transformers in NLP, ViT [21], as a pioneering work, introduced vision Transformers into computer vision tasks. Specifically, toward image classification, ViT decomposes each image into a sequence of patch embeddings and exploits multiple Transformer layers to encode them into Transformer tokens, thereby achieving state-of-the-art classification accuracy. DeiT [33] further delved into data-efficient training and distillation mechanisms to alleviate dependence of vision Transformers on extremely large-scale training data, e.g., using JFT-300M in ViT. For more complicated high-level image analysis tasks, including object detection, a considerable number of variants of vision Transformer [22], [23], [24], [34], [35], [36] have been developed, all of which consistently outperform CNN-based models. For instance, PVT [23] incorporates an effective multiscale Transformer architecture to generate discriminative hierarchical features by imitating the classic feature pyramid structure in CNNs, naturally favoring dense predictions. However, a prominent problem still remains in these concurrent works, which is the heavy computational complexity brought by the global self-attention computation seriously hinders the practice of vision Transformers, especially on high-resolution images.

In the context of remote sensing, vision Transformers have also received increasing attention in the past two years. For example, some Transformer-based change detection methods [43], [72], [73], [74] have been explored that attain impressive superiority over CNN-based algorithms. Chen et al. [43] proposed a bitemporal image Transformer, termed BIT, for remote sensing image change detection, which established spatial–temporal contextual relations with a suppressed number of semantic tokens to improve model efficiency. Despite the above-mentioned progress, only a few remote sensing object detectors based on vision Transformers [46], [75] are available in this field. In [75], a geospatial Transformer with two auxiliary geospatial attention modules was implemented on synthetic aperture radar (SAR) images, yet only for aircraft detection. Hence, there is an extreme scarcity of Transformer-based object detectors with low computational costs and high efficiency for very high-resolution optical remote sensing images.

III. Methodology

The aforementioned three obstacles, i.e., high computational costs, lack of inductive bias, and arbitrary orientation, significantly hinder the practical potential and detection performance of vision Transformers in remote sensing scenarios. Thus, we propose an advanced framework in this work, namely, EIA-PVT, to address these three challenges and improve both the efficiency and accuracy of Transformer-based object detectors. In this section, we will interpret how to jointly overcome the three obstacles by using a unified and concise pipeline. Specifically, the proposed framework and its workflow will be first overviewed, and then, three crucial techniques or strate-

gies, i.e., adaptive multigrained routing, inductive dual-path encoding, and angle tokenization, will be presented in detail.

A. Overview

Geospatial objects in remote sensing imagery are usually characterized by their large variations in scale, making the detection much more difficult than nature-scene objects. A typical and effective solution to this issue is to construct CNN-based multiscale features for different remote sensing objects [5], [77], [78]. Similarly, Transformer-based detectors should be able to generate multiscale representations stage by stage. As shown in Fig. 3, our proposed EIA-PVT is composed of four successive stages for generating features of different scales. Given an input image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, the four hierarchical representations can be derived, denoted as $\{\mathcal{F}_i\}$, with different feature dimensions C_i $(i = 1, 2, 3, 4)$, where i represents the stage index. The corresponding scales of the four stages are $(H/4) \times (W/4)$, $(H/8) \times (W/8)$, $(H/16) \times (W/16)$, and $(H/32) \times (W/32)$, respectively. These features of four different scales will be utilized in the corresponding external detection heads to produce the final detection results, in the form of oriented bounding boxes (x, y, w, h, θ). Each stage contains three pivotal components, i.e., PPE, EIE, and ATM, which make the proposed EIA-PVT superior to current Transformers. PPE aims to flexibly control the scale of the hierarchical representations and establish a robust scale-invariant feature pyramid for subsequent detection. EIE, which is the core component of EIA-PVT, mainly contains two modules, AMGR and IDPM, and is designated to efficiently and inductively encode raw patches into discriminative representations, with reduced spatial redundancy and richer semantic information. It is worth noting that the number of stacked EIEs in stage i, denoted as N_i, varies with stages to shape Transformer models with different sizes. The last component, ATM, is proposed to effectuate angle modeling and learning, specifically for Transformer-based oriented object detectors.

Fig. 4 depicts the streamlined workflow of the ith stage in EIA-PVT, showing how the three components operate and interact in a synergistic and harmonious manner.

1) PPE: At the beginning of the ith stage, the feature map from the $(i-1)$st stage, denoted as $\mathcal{F}_{i-1} \in \mathbb{R}^{H_{i-1}\times W_{i-1}\times C_{i-1}}$, is fed into PPE to construct a patch sequence of dimension $(H_{i-1}W_{i-1}/\mathcal{P}_i^2) \times C_i$, where H_{i-1} and W_{i-1} are the height and width of the feature map, respectively, and $\mathcal{P}_i$ represents the spatial dimension reduction ratio. Unlike the vanilla PVT, which simply splits and flattens 3-D features into patches via a linear patch embedding layer, our PPE includes a pyramid-projection layer to introduce the inductive bias with respect to scale invariance in the 3-D structures, which alleviates the information loss caused by token merging in EIE. As depicted in Fig. 4, $\mathcal{D}_i = \{d_{i,k}\}$, where $k = 1, 2, \ldots, K$, represents the set of dilation rates of the pyramid-projection layer with K levels. Multiple dilated convolutions are applied to the input feature $\mathcal{F}_{i-1}$, and the outputs are concatenated to generate a new 3-D feature $\mathcal{F}'_i \in \mathbb{R}^{(H_{i-1}/\mathcal{P}_i)\times(W_{i-1}/\mathcal{P}_i)\times C_i}$, formulated as follows:

$$\mathcal{F}'_i = \text{Concat}([\text{DConv}_{i,k}(\mathcal{F}_{i-1}; d_{i,k}, \mathcal{P}_i) \,|\, d_{i,k} \in \mathcal{D}_i]) \tag{1}$$

where $\text{DConv}_{i,k}(\cdot)$ represents the kth convolutional projection based on the corresponding dilation rate $d_{i,k}$, while stride convolutions are employed to reduce the spatial dimension according to the preset $\mathcal{P}_i$. Concat([;]) represents the feature concatenation function. In addition, different stages adopt different dilation rate sets $\mathcal{D}_i$, and the number of dilated convolutions $|\mathcal{D}_i|$ also varies accordingly, which will be discussed in Section IV. Next, the concatenated feature $\mathcal{F}'_i$ is processed by the Gaussian error linear unit (GeLU) activation function, followed by a layer normalization (LN) function. The final output is evenly divided into patch embeddings, obligated in most Transformers, formulated as follows:

$$\mathbf{x}_i \in \mathbb{R}^{H_iW_i\times C_i} \leftarrow \text{LayerNorm}(\text{GeLU}(\mathcal{F}'_i)). \tag{2}$$

In this way, PPE introduces scale invariance as an inductive bias via multilevel dilated convolutions and also extracts multiscale contexts to construct a feature pyramid.

2) EIE: Similar to the design of convolutional layers, there are N_i stacked EIEs in EIA-PVT. For the sake of simplicity, we only consider one EIE in the following descriptions and formulations, unless otherwise specified. The goal of EIE is to efficiently encode input patch embeddings $\mathbf{x}_i$ into inductive tokens $\mathbf{f}'_i$, equipped with appropriate inductive bias and sparsity. As illustrated in Fig. 4, EIE consists of two main modules, AMGR and IDPM, where the former adaptively and dynamically generates multigrained patches with suppressed spatial redundancy, and the latter serves to model both global dependencies and localities. Notably, compared with the original tokens in the vanilla Transformer encoders, these tokens derived by AMGR and passed in EIE are sparse, significantly reducing the computational complexity in IDPM, especially for self-attention calculations. Specifically, given an input sequence of patches, AMGR can work in parallel on nonoverlapping feature regions $\mathbf{r}_i$ and adaptively route different granularities to generate multigrained patches $\tilde{r}_i$. This routing mechanism is adaptive and data dependent. For example, as can be seen in Fig. 4, the foreground regions of the baseball field (BF) are assigned to be finer-grained (more tokens within a fixed-size window) than other background regions to reduce spatial redundancy. Then, a pooling operation is employed to spatially average these multigrained patches into the same scale to form sparse tokens $\tilde{t}_i$, as follows:

$$\begin{aligned} &\tilde{t}_i = \text{Pooling}(\tilde{r}_i) \in \mathbb{R}^{\mathcal{S}_{i,n}\times C_i}, \quad \tilde{r}_i = \text{AMGR}(\mathbf{r}_i) \\ &\text{s.t.} \quad \mathcal{S}_{i,n} \leqslant H_iW_i, \quad n \in \{1, 2, \ldots, N_i\} \end{aligned} \tag{3}$$

where $\mathcal{S}_{i,n}$ refers to the total number of tokens in the nth EIE of the ith stage. Generally, the sparse tokens are flexibly compatible with most Transformer encoders, while our proposed IDPM is proficient in modulating them both globally and locally into more discriminative tokens. Finally, an unpooling operation is exploited to restore these encoded tokens to the original resolution, termed inductive tokens

$$\begin{aligned} &\mathbf{f}'_i = \mathbf{x}_i \oplus \text{Unpooling}(\hat{\mathbf{t}}_i) \in \mathbb{R}^{H_iW_i\times C_i} \\ &\text{where} \quad \hat{\mathbf{t}}_i \leftarrow \mathbf{t}_i = \text{IDPM}(\tilde{t}_i, \mathbf{x}_i) \end{aligned} \tag{4}$$

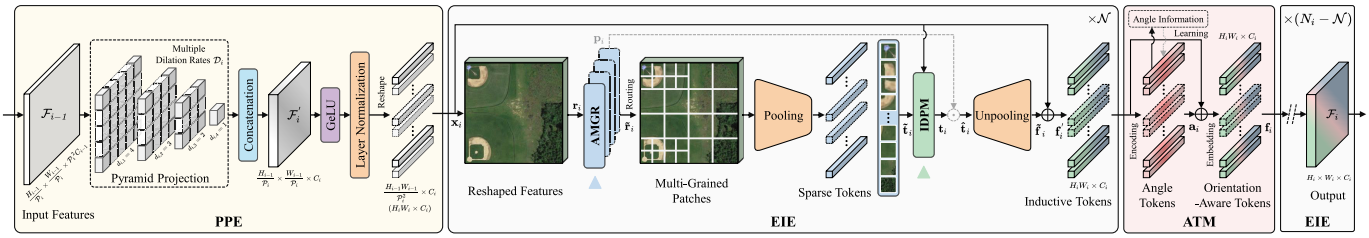


Fig. 4. Workflow of the ith stage in the proposed EIA-PVT, in which the features from the previous stage (stage $i-1$), is processed by PPE, EIE, and ATM to enhance the token semantics and then shaped into the representations with the required scale in the next stage (stage $i+1$). EIE can efficiently perform long-range contextual encoding at low computational complexity and cooperate with PPE in complementing inductive bias, while ATM explicitly models angle information for accurately oriented object detection in remote sensing images. The gray dash lines are only activated in the training phase, and the residual connections are added to typically mitigate the vanishing gradient problem and boost feature details [76]. The total number of EIEs in stage i is N_i, while ATM is inserted after the first $\mathcal{N}$ EIEs, followed by the remaining $(N_i - \mathcal{N})$ EIEs.

and $\oplus$ denotes elementwise addition. Thanks to the token sparsity handled by AMGR, the calculations in EIE are more efficient with lower computational costs than the vanilla Transformer encoders. Furthermore, IDPM introduces a dual-path encoding, where local and global modeling are incorporated to develop distinctive tokens with powerful induction and semantics. More details about AMGR and IDPM will be provided in Section III-B and Section III-C, respectively.

3) ATM: Following the design in vision Transformers [21], [22], [23], each stage usually consists of multiple consecutive encoders, i.e., EIEs in our framework. However, different from previous Transformers, toward accurately detecting oriented objects in remote sensing images, our EIA-PVT contains an extra ATM to explicitly tokenize angle information, thus alleviating the problem of hard-to-learn orientation knowledge in Transformer-based detectors. As illustrated in Fig. 4, for generality and convenience, the ATM is designated and inserted after the first $\mathcal{N}$ EIEs, followed by the remaining $(N_i - \mathcal{N})$ EIEs. Moreover, ATM properly bridges the detection head enclosing orientation supervision and the intermediate tokens to facilitate angle learning, without affecting the EIE architecture. These orientation-aware tokens enhanced by ATM, denoted as $\mathbf{f}_i$, can be directly fed into the following EIEs, while the final feature $\mathcal{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ generated by the last EIE will be sent to the next stage. More details about ATM will be introduced in Section III-D.

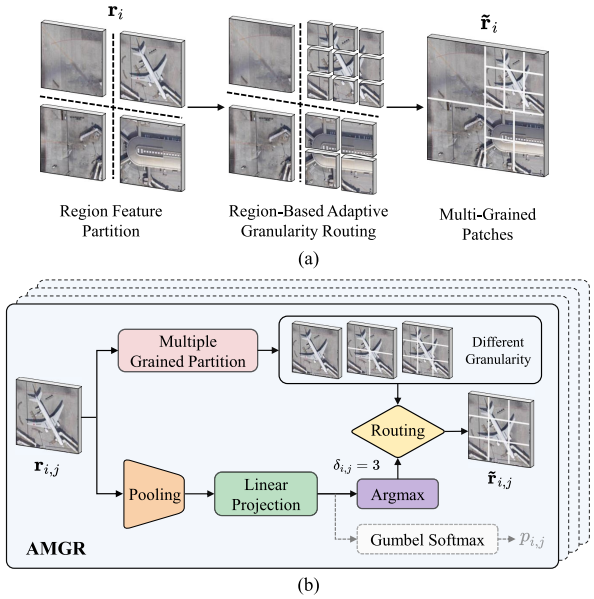


Fig. 5. Illustration of the adaptive multigrained routing mechanism in AMGR. (a) Input feature is split into multiple nonoverlapping regular regions to perform region-based routing in parallel, and the results are concatenated into the multigrained patches. (b) For each region feature, multiple groups of patches are generated according to different granularities, one of which will be chosen by the routing network. The Gumbel softmax with the dash lines, are only activated and utilized for training.

B. Adaptive Multigrained Routing

1) Motivations: The proposed AMGR introduces a novel multigrained routing mechanism, aimed at adaptively nominating diverse granularity for different image or feature patches, thereby reducing computational cost but only slightly or without degrading detection performance. As investigated in Fig. 2, in the context of remote sensing object detection, neighboring tokens in Transformer encoders typically have strong correlations, indicating the significant potential of diminishing spatial redundancy via patchy sparsity for efficient computation. This sparsity strategy is dynamically adaptive to input images, i.e., data-dependent, such that it is feasible to maintain detection accuracy. For example, only informative tokens related to the foreground of an image or with strong semantics are designated to be fine-grained, while those uninformative tokens, containing only background without semantics, should be coarse-grained to save computational resources [27]. For clear understanding, Algorithm 1 concisely characterizes the proposed adaptive multigrained routing mechanism and its main steps.

2) Architectures: To facilitate mixed-grained patches in space, as shown in Fig. 5(a), the input feature is first partitioned into multiple regular regions $\mathbf{r}_i$ based on nonoverlapping windows of size $\mathcal{M} \times \mathcal{M}$. Accordingly, granularity routing can be performed on these windows in parallel to improve efficiency. Some irregular region partitioning strategies, such as SLIC superpixels [79], may also work, but this will introduce additional computation that is inefficient and unfriendly, contrary to our purpose. Then, a set of granularity candidates, denoted as $\mathcal{G} = \{g_1, g_2, \ldots, g_L\}$, is defined to provide patch size options for each feature region, where L indicates the number of granularity candidates and the granularity g represents the side length of a patch. For example,

Algorithm 1 Adaptive Multigrained Routing

Input: A region patch $\mathbf{r}_{i,j}$

Output: The multigrained region patch $\tilde{r}_{i,j}$

Initialization:

1: Define a granularity candidate set $\mathcal{G} = \{g_1, g_2, \ldots, g_L\}$
2: Generate multiple groups of region patches corresponding to different granularity from $\mathcal{G}$

Inference Phase:

1: Project $\mathbf{r}_{i,j}$ to the granularity logits $\mu_{i,j}$ by Eq. (5)
2: Determine the granularity index $\delta_{i,j}$ for $\mathbf{r}_{i,j}$ by Eq. (6)
3: Route a specific group with the granularity $g_{\delta_{i,j}}$ according to the index $\delta_{i,j}$ to represent $\tilde{r}_{i,j}$
4: Pool the multigrained region patch $\tilde{r}_{i,j}$ to form sparse tokens $\tilde{t}_{i,j}$ by Eq. (7)
5: Concatenate all these sparse tokens to form $\tilde{t}_i$ for Eq. (4)

Training Phase:

1: Project $\mathbf{r}_{i,j}$ to the granularity logits $\mu_{i,j}$ by Eq. (5)
2: Determine the granularity index $\delta_{i,j}$ by Eq. (8)
3: Route the granularity $g_{\delta_{i,j}}$ for $\tilde{r}_{i,j}$ according to $\delta_{i,j}$
4: Pool $\tilde{r}_{i,j}$ to form sparse tokens $\tilde{t}_{i,j}$ by Eq. (7)
5: Calculate the routing score $p_{i,j}$ by Eq. (9)
6: Generate inductive tokens $\mathbf{t}_{i,j}$ and $\hat{t}_{i,j}$ by Eq. (10)
7: Calculate the training loss by Eq. (11) and Eq. (12)

$g = 4$ should typically correspond to a 4×4 patch. Since all the mixed-grained patches will be pooled into tokens or queries with the same dimensions for encoding, a larger g generally means a larger sized and coarser-grained feature patch, i.e., sparser tokens and less computation in EIE. Therefore, the region size or window size can be set to the maximum granularity in practice, formulated as $\mathcal{M} = \max(\mathcal{G})$. Given a region feature $\mathbf{r}_{i,j} \subseteq \mathbf{r}_i$, where $j \in \{1, 2, \ldots, \lceil (H_i/\mathcal{M}) \rceil \cdot \lceil (W_i/\mathcal{M}) \rceil\}$ is the region index, the specific granularity should be adaptively allocated from $\mathcal{G}$ via a compact routing network. Specifically, as illustrated in Fig. 5(b), multiple groups of region patches in terms of L different granularity are first generated to be selected by routing. Then, an average pooling operation followed by a linear projection is employed to map the raw region feature $\mathbf{r}_{i,j}$ to a representation of length L, each element indicating the corresponding granularity logits

$$\mu_{i,j} = \frac{1}{\mathcal{M}^2} \sum_{m=1}^{\mathcal{M}^2} \mathbf{r}_{i,j}^{(m)} \cdot \boldsymbol{\omega}_i + \beta_i \tag{5}$$

where $\boldsymbol{\omega}_i \in \mathbb{R}^{C_i \times L}$ and $\beta_i \in \mathbb{R}^{1 \times L}$ denote the weights and bias, respectively. Next, the granularity indices can be derived based on the argument-maximum indexing function $\arg\max(\cdot)$ by

$$\delta_{i,j} = \arg\max_l (\mu_{i,j}^{(l)}) \in \{1, 2, \ldots, L\}. \tag{6}$$

According to this predicted index, one group of region patches with the granularity $g_{\delta_{i,j}}$ is selected by routing, denoted as $\tilde{r}_{i,j} \in \mathbb{R}^{\Omega_{i,j} \times g_{\delta_{i,j}}^2 \times C_i}$, where $\Omega_{i,j} = \lceil (\mathcal{M}/g_{\delta_{i,j}}) \rceil^2$ indicates the number of patches in this group. The patch groups of all regions are collected to form the final multigrained patches $\tilde{r}_i$ as the output of AMGR. As depicted in Fig. 4 and introduced in Section III-A, the spatial mean vector of each element in $\tilde{r}_i$ from AMGR will be molded to the same scale by the pooling operation, serving as the representative token to compose the sparse tokens $\tilde{t}_i \in \mathbb{R}^{\mathcal{S}_i \times C_i}$ [n in (3) is omitted for brevity]

$$\tilde{t}_{i,j} = \frac{1}{g_{\delta_{i,j}}^2} \sum_{m=1}^{g_{\delta_{i,j}}^2} \tilde{r}_{i,j}^{(m)} \in \mathbb{R}^{\Omega_{i,j} \times C_i}$$
$$\text{s.t.} \quad \mathcal{S}_i \equiv \sum_{j=1}^{\lceil \frac{H_i}{\mathcal{M}} \rceil \cdot \lceil \frac{W_i}{\mathcal{M}} \rceil} \Omega_{i,j}. \tag{7}$$

Consequently, the number of tokens in the jth feature region is reduced to $(1/g_{\delta_{i,j}}^2)$ of the previous, and these sparse tokens will be fed into IDPM as sparse queries to replace the original nonsparse queries for better encoding efficiency.

3) End-to-End Optimization: The proposed AMGR is advantageous to static token fusion strategies, due to its dynamic adaptability to various input images, i.e., data dependent, which should be end-to-end trainable. To this end, the discrete determined decisions in AMGR, i.e., (6), are replaced by stochastic sampling [80], [81], [82]. With a sequence of unnormalized log probabilities subject to a specific categorical distribution, during training, the discrete granularity indices are determined by

$$\delta_{i,j} = \arg\max_l (\mu_{i,j}^{(l)} + \gamma^{(l)})$$
$$\text{s.t.} \quad \gamma^{(l)} \sim \Gamma(0, 1) \tag{8}$$

where $\gamma^{(l)}$ represents the Gumbel noise sampled from the standard Gumbel distribution $\Gamma(0, 1)$. However, the above-mentioned formulation involves a hard decision process, which is still nondifferentiable for training. To handle this issue, the Gumbel-softmax relaxation [83] is introduced to provide a continuously differentiable approximation, and then, the routing score can be predicted from the granularity index, as follows:

$$p_{i,j} = \text{softmax}((\mu_{i,j}^{(\delta_{i,j})} + \gamma^{(\delta_{i,j})})/\tau)$$
$$= \frac{\exp((\mu_{i,j}^{(\delta_{i,j})} + \gamma^{(\delta_{i,j})})/\tau))}{\sum_{l=1}^{L} \exp((\mu_{i,j}^{(l)} + \gamma^{(l)})/\tau))} \in [0, 1] \tag{9}$$

where τ is a temperature parameter. Furthermore, we exploit the straight-through estimator to compute the gradients of granularity logits during the backward pass

$$\hat{t}_{i,j} = \begin{cases} \mathbf{t}_{i,j}, & \text{for forward pass} \\ p_{i,j} \cdot \mathbf{t}_{i,j}, & \text{for backward pass.} \end{cases} \tag{10}$$

It can be observed that this estimator is biased due to the misalignment between the forward and backward passes, but it is empirically effective, as verified in [80] and [81]. Remarkably, the above-mentioned stochastic process is only aimed at facilitating end-to-end training, while the inference phase is free from the stochastic sampling and the exponential calculations in (8) and (9), respectively, to improve efficiency and usability.

4) Loss Function: The above-mentioned design can solve the nondifferentiability problem of backpropagation in AMGR. Nevertheless, there is still a mismatch between the network update target (i.e., determined by the training loss) and the role of AMGR. In other words, despite multiple granularity options, AMGR tends to route finer grained representations or allocate more tokens for each region, which may generally benefit the detection performance at the expense of token sparsity. To overcome this problem, we propose a computational budget constraint strategy to strictly guide the entire framework toward a decent tradeoff between high accuracy and low complexity during training. Concretely, considering that the main computational load of the proposed EIA-PVT comes from EIEs, especially IDPM, we define the computational complexity associated with a token in the nth EIE of the ith stage as $\mathcal{C}_{i,n}$. Then, the computational-complexity ratio of the sparse tokens to the original nonsparse tokens is computed as follows:

$$\eta = \frac{\sum_{i=1}^{4}\sum_{n=1}^{N_i}(\mathcal{C}_{i,n}\cdot\mathcal{S}_{i,n})}{\sum_{i=1}^{4}\sum_{n=1}^{N_i}(\mathcal{C}_{i,n}\cdot H_i W_i)} \tag{11}$$

where $\mathcal{S}_{i,n}$ represents the number of tokens, as in (3). We further define a computational budget term, denoted as $\sigma \in [0, 1]$, to indicate the suggested computational-complexity ratio for η. Therefore, we introduce it into the training loss, and the total training loss can be refined as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{det}} + \alpha\mathcal{L}_{\text{comp}} \\ &\text{where}\quad \mathcal{L}_{\text{comp}} = (\eta-\sigma)^2\end{aligned} \tag{12}$$

where $\mathcal{L}_{\text{det}}$ represents the remote sensing object detection loss. $\mathcal{L}_{\text{comp}}$ constrains the computational cost by minimizing the difference between the complexity ratio η and its targeted bound σ. Additionally, α is utilized to control the relative importance of the different losses. This computational budget constraint strategy provides EIEs with a complexity estimation, properly balancing efficiency and performance.

C. Inductive Dual-Path Encoding

1) Motivations: A vanilla vision Transformer encoder [21], [22], [23] consists of three major components, multihead self-attention (MHSA), the feed-forward network, and LN. The self-attention mechanism essentially facilitates modeling long-distance feature dependencies, which is desirable for remote sensing object detection. Unfortunately, in spite of performance gains, constructing such global representations comes at the cost of deteriorating powerful local correlations and details, compromising the discriminability between foregrounds and backgrounds [31], [32]. On the other hand, it has been concluded [25], [30] that explicitly introducing inductive bias into Transformers can significantly enhance the sample efficiency and parameter efficiency, thereby getting rid of the requirement for data-plentiful training regimes, and the locality of convolutions is able to behave as representative inductive bias. Actually, such convolutional inductive bias is especially valuable and friendly for the task of remote sensing object detection, due to the unavailability of large-scale fully annotated training samples. To sum up, the performance and generalizability of Transformer-based detectors can be further explored and improved by efficiently encoding appropriate inductive bias. Therefore, the above-mentioned facts motivate the proposal of IDPM to replace the vanilla Transformer encoder to handle inductive bias. Notably, as described in Section III-A, the proposed PPE has epitomized inductive bias from the perspective of scale invariance via convolutional pyramid projections, based on four hierarchical stages only. Complementary to PPE, IDPM aims to further flexibly enhance intrinsic inductive bias for each encoder, typically locality, which concurrently combines convolutional and self-attention layers in a compact and plug-and-play architecture.

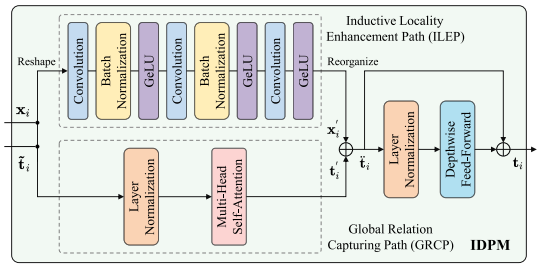


Fig. 6. Structure of the proposed IDPM, which mainly includes two parallel encoding paths for inductive locality enhancement and global relation capturing, respectively. ⊕ indicates the elementwise addition. With appropriate inductive bias, the final generated inductive tokens contain both advanced local details and global long-range dependencies.

2) Architectures: As illustrated in Fig. 6, two parallel paths are devised in the proposed IDPM, i.e., the inductive locality enhancement path (ILEP) and the global relation capturing path (GRCP), where the former introduces inductive bias and interacts with the latter to reinforce both local and global representations. Essentially, GRCP is still the main path for information encoding, generally following the design of vanilla Transformer encoders, while ILEP serves as an auxiliary path, responsible for modulating decent inductive bias to be fused in the tokens. Specifically, as shown in Fig. 6, ILEP is composed of three stacked convolutional layers, interleaved with batch normalization and GeLU activation layers. Given the original nonsparse token $\mathbf{x}_i$ as input, which is first reshaped as a convolution-oriented feature with a 3-D structure, ILEP then translates it into the inductive feature with the rich convolutional locality. Since computing convolutions is more efficient than self-attention, it is reasonable and convenient to directly operate on nonsparse 3-D features without any local spatial information loss. Finally, according to the sparse token $\tilde{t}_i$ from AMGR, which should also be fed into IDPM, the features processed by ILEP are pooled and reorganized into the identical structure as $\tilde{t}_i$, for subsequent elementwise modulation. The procedure of ILEP can be described as follows:

$$\mathbf{x}_i' = \text{ILEP}(\mathbf{x}_i, \tilde{t}_i) \tag{13}$$

where $\mathbf{x}_i'$ contains enhanced locality as inductive bias. To capture the global context, the other path, i.e., GRCP, follows the general design of PVT encoders [23], yet distinctively taking both $\tilde{t}_i$ and $\mathbf{x}_i$ as input. The sparse token $\tilde{t}_i$ will be

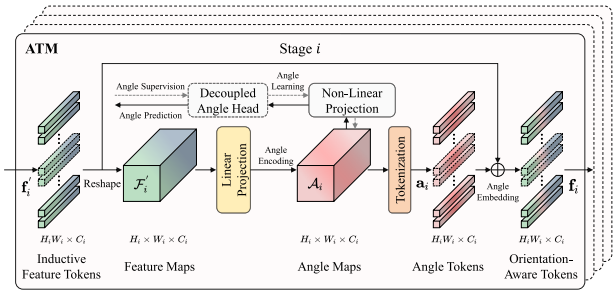


Fig. 7. Illustration of the proposed ATM. There are four ATMs in EIA-PVT, each of which corresponds to one of four hierarchical stages with a specific feature level. In stage i, under the supervision of the decoupled angle head, orientation knowledge is encoded and tokenized into angle tokens, which facilitates elementwise angle embedding for the original inductive feature tokens to generate the final orientation-aware tokens as output.

exploited to generate sparse queries in self-attention, while $\mathbf{x}_i$ is exploited for keys and values. Sparse queries result in dramatically reduced computational complexity, as described in Sections III-A and III-B. Specifically, in Fig. 6, $\tilde{t}_i$ and $\mathbf{x}_i$ are dealt with by LN and MHSA to produce the global representation $\mathbf{t}'_i$ as follows:

$$\begin{aligned}\mathbf{t}'_i &= \mathrm{GRCP}(\mathbf{x}_i, \tilde{t}_i) \\ &= \mathrm{MHSA}\big(\mathrm{LN}(\tilde{t}_i)W_i^{\mathcal{Q}}, \mathrm{LN}(\mathbf{x}_i)W_i^{\mathcal{K}}, \mathrm{LN}(\mathbf{x}_i)W_i^{\mathcal{V}}\big)\end{aligned} \tag{14}$$

where $W_i^{\mathcal{Q}}$, $W_i^{\mathcal{K}}$, and $W_i^{\mathcal{V}}$ refer to linear projection matrices that project $\tilde{t}_i$ or $\mathbf{x}_i$ into queries, keys, and values, respectively. Therefore, compared with vanilla Transformer encoders, the self-attention mechanism in our GRCP requires much less computation, due to token sparsity. Next, the outputs of ILEP and GRCP are aggregated to yield an improved representation, formulated as

$$\ddot{\mathbf{t}}_i = \mathbf{t}'_i \oplus \mathbf{x}'_i \tag{15}$$

where $\ddot{\mathbf{t}}_i$ is the modulated token carrying both global relations and locality inductive bias. As shown in Fig. 6, we place LN and a depthwise feed-forward network (DW-FFN) with a residual connection on the top of IDPM to generate the final inductive token $\mathbf{t}_i$, as follows:

$$\mathbf{t}_i = \ddot{\mathbf{t}}_i \oplus \mathrm{DW\text{-}FFN}(\mathrm{LN}(\ddot{\mathbf{t}}_i)). \tag{16}$$

It is noteworthy that we replace the original FFN with our DW-FFN based on depthwise convolutions [84] to further boost the inductive bias and guarantee information exchange within the inductive token [24].

D. Angle Tokenization

1) Motivations: Angle prediction for oriented object detection in remote sensing images has always been a critical issue, as small angular deviations may lead to large discrepancies in rotated bounding-box regression, and then degrade the final detection accuracy [54]. To tackle this challenge, previous CNN-based oriented object detectors tend to develop additional rotation refinement modules to boost the angle representation ability [1], [50], [52], [68], [85]. Nevertheless, these CNN-specific angle learning strategies are unsuitable or suboptimal for Transformer-based oriented object detectors, in terms of their different information encoding schemes. Typically, the regular global operations of self-attention essentially limit encoding generalization to rotation variations of remote sensing objects. It is intractable to force Transformer encoders to implicitly learn diverse object direction knowledge only by simply introducing an extra angle representation, like those CNN-based detectors. Therefore, an urgent problem to be solved is how to smoothly adapt Transformer-style architectures to oriented object detection without major modifications to the core Transformer encoders (i.e., EIEs in this work) and heavy computational overhead. Motivated by this, we demonstrate that an effective solution is to bridge these encoders with the angle prediction branch to explicitly encode accurate object orientation information into the tokens passed into the whole workflow. To this end, we propose a lightweight and flexibly pluggable module, called ATM, which enables efficient and robust angle learning, especially for Transformer-based detectors against various rotation disturbances.

2) Design Details: From a high-level perspective, Fig. 4 presents the role and usage of the proposed ATM in the whole workflow of EIA-PVT. Each stage contains an ATM that can be conveniently inserted into consecutive EIEs, with one end connected to the inductive tokens generated by the previous EIE, and the other end converting the output orientation-aware tokens to the next EIE. The ATM aims to discriminate feature tokens with appropriate object orientation knowledge by tokenizing angle information. Hence, intuitively, it should be inserted between EIEs, rather than before or after; otherwise, it will lead to either indiscriminative angle representations or insufficient information exchange between feature tokens and angle tokens. For instance, at the ith stage, the ATM is placed after $\mathcal{N}$ EIEs, followed by $(N_i - \mathcal{N})$ EIEs, while $\mathcal{N}$ is set to 1 by default. Technically, ATM provides a novel Transformer-friendly paradigm of angle encoding, embedding, and learning, for rotated remote sensing objects. Fig. 7 presents the structural details of ATM. With inductive tokens as input, ATM encodes orientation information as angle tokens under the angle supervision from the detection head. The angle tokens are then embedded back into the original inductive tokens to yield orientation-aware tokens as output. Concretely, as depicted in Fig. 7, taking stage i as an example, given an inductive feature token $\mathbf{f}'_i \in \mathbb{R}^{H_i W_i \times C_i}$ [as in (4)], ATM first reshapes it into a 3-D feature map $\mathcal{F}'_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ with spatial context, and then linearly projects $\mathcal{F}'_i$ to an angle map $\mathcal{A}_i$, as follows:

$$\mathcal{A}_i = \mathrm{LP}(\mathcal{F}'_i) \in \mathbb{R}^{H_i \times W_i \times C_i} \tag{17}$$

where $\mathrm{LP}(\cdot)$ represents a learnable linear projection function, such as convolutional or fully connected layers. More importantly, to activate the capability of $\mathcal{A}_i$ to iteratively learn and encode orientation knowledge, we devise an explicit connection between the angle map and its corresponding angle head via a nonlinear projection. This angle head serves as a branch of the external detection head; yet is decoupled from the regression head and the classification head to separately provide accurate angle supervision and avoid supervision distraction. Furthermore, this design enables ATM to be trained end-to-end. As shown in Fig. 7, conditioned by

the above-mentioned orientation information, the angle map can be further tokenized into an angle token $\mathbf{a}_i \in \mathbb{R}^{H_i W_i \times C_i}$, with the same dimensions as $\mathbf{f}_i'$. Then, angle embedding is performed based on the original inductive token and the newly derived angle token, as follows:

$$\mathbf{f}_i = \mathbf{f}_i' \oplus \mathbf{a}_i \in \mathbb{R}^{H_i W_i \times C_i} \tag{18}$$

$\mathbf{f}_i$ is named an orientation-aware token, in which the discriminative angle information has been incorporated and enhanced for the repeated encoding process in the subsequent EIEs.

3) Loss Function: As presented in Fig. 3, the proposed ATM explicitly bridges the Transformer encoders and the detection head, without adverse effect on the EIE architecture and the end-to-end training of EIA-PVT. Following the representations in [69] and [54], given a predicted oriented bounding box $\mathbf{b} = (x, y, w, h, \theta)$ and its ground truth $\mathbf{b}^* = (x^*, y^*, w^*, h^*, \theta^*)$, where θ and θ^* denote the acute angle with the x-axis, the detection loss function in (12) is defined as

$$\mathcal{L}_{\text{det}} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{loc}(\mathbf{b}, \mathbf{b}^*) + \lambda_3 \mathcal{L}_{\text{angle}}(\theta, \theta^*). \tag{19}$$

The focal loss [14] and the approximate SkewIoU loss [69] are employed as the classification loss $\mathcal{L}_{\text{cls}}$ and the localization loss $\mathcal{L}_{loc}$, respectively. To supervise angle prediction, the smooth-L1 loss is exploited for $\mathcal{L}_{\text{angle}}$ as follows:

$$\begin{aligned}\mathcal{L}_{\text{angle}}(\theta, \theta^*) &= \text{Smooth-L1}(\theta - \theta^*) \\ &= \begin{cases} 0.5(\theta - \theta^*)^2, & \text{if } |\theta - \theta^*| < 1 \\ |\theta - \theta^*| - 0.5, & \text{otherwise} \end{cases}\end{aligned} \tag{20}$$

where λ_1, λ_2, and λ_3 are hyperparameters that control the tradeoff and are set to 1 by default.

IV. Experiments and Analysis

In this section, extensive experiments on oriented object detection in remote sensing images were conducted to evaluate the effectiveness and superiority of the proposed EIA-PVT. We first briefly introduce the datasets and evaluation metrics, and then, comprehensively present and analyze all experimental results, including ablation studies and comparisons with state-of-the-art methods.

A. Dataset Description

To comprehensively evaluate the proposed EIA-PVT framework, two publicly available datasets for oriented object detection in remote sensing images, namely, DOTA [47] and DIOR-R [50], are utilized in our experiments.

1) DOTA Dataset: DOTA [47] is the most representative large-scale oriented object detection dataset, consisting of 2806 remote sensing images and 188 282 instances, categorizing into 15 geospatial classes, defined as plane (PL), baseball diamond (BD), bridge (BR), ground field track (GFT), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). Notably, this dataset contains a variety of complex and challenging remote sensing scenarios, such as considerable variances in object scale, shape, and aspect ratios. Furthermore, most annotated geospatial objects are densely packed and cluttered with arbitrary orientations under changeable imaging conditions. Considering its difficulty, we carry out ablation studies on this dataset to verify the effectiveness of our method. In the experiments, 1/2, 1/6, and 1/3 of the original images are randomly selected for training, validation, and testing, respectively. Additionally, since the image resolutions range extremely from less than 800×800 pixels to more than 4000×4000 pixels, all images are cropped into patches of size 640×640 pixels, with an overlap of 150 pixels for efficient training.

2) DIOR-R Dataset: DIOR-R [50] is currently the largest dataset for oriented object detection in remote sensing scenarios, consisting of 23 463 images with a total of 190 288 instances. This dataset covers 20 common geospatial object categories, APL, airport (APO), BF, BC, BR, chimney (CH), dam (DAM), expressway toll station (ETS), expressway service area (ESA), golf field (GF), ground track field (GTF), HA, overpass (OP), SH, stadium (STA), ST, TC, train station (TS), vehicle (VE), and windmill (WM). To facilitate performance evaluation and comparison, this dataset is randomly partitioned into two halves to form the training set and test set in the experiments.

B. Evaluation Metrics and Implementation Setup

Average precision (AP) is utilized as the main metric to quantitatively evaluate the detection performance of diverse detectors. Specifically, we divide the detection results into four different cases, i.e., true positive (TP), false positive (FP), true negative (TN), and false negative (FN), and then count the number of samples falling into each case to calculate the precision rate (P) and recall rate (R), defined as follows:

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{21}$$

$$\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{22}$$

By applying different thresholds to measure multiple precision and recall values, the precision-recall (PR) curve can be plotted. AP can be expressed as $\text{AP} = \int_0^1 \text{P(R)} d\text{R}$, i.e., the area under the PR curve. Generally, an AP is measured for each object class, and the mAP represents the mean of all class APs, computed by

$$\text{mAP} = \frac{1}{\mathcal{C}} \sum_{c=1}^{\mathcal{C}} \text{AP}_c \tag{23}$$

where $\mathcal{C}$ denotes the total number of object categories in the dataset, and higher mAP typically indicates higher detection accuracy. In addition, we adopt the number of parameters (hereafter abbreviated as Param), floating point operations (FLOPs), and latency to represent the computational complexity and efficiency of the different models in the experiments, especially considering that this work aims to improve both the efficiency and accuracy of Transformers for remote sensing object detection in practice. For fair comparisons, unless otherwise specified, we only investigate the Param and FLOPs of the backbones of all detectors.

TABLE I

PERFORMANCE EVALUATION AND COMPARISON OF THE BASELINE MODEL, THE PROPOSED EIA-PVT, AND ITS THREE REDUCED MODELS

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP (Δ)	Param (M)	FLOPs (G)	Latency (ms)
Baseline	89.4	69.1	31.7	57.7	60.4	76.0	80.2	90.5	60.9	84.9	64.0	54.7	56.6	50.8	58.2	65.7	**24.9**	41.3	78
E-PVT	89.0	63.5	34.8	52.8	56.9	74.5	77.8	90.8	66.0	84.8	63.9	61.9	63.1	50.8	55.8	65.8 (+0.1)	24.9	**30.4**	**67**
I-PVT	**90.0**	71.5	37.5	**67.5**	59.3	78.5	79.7	**90.9**	65.0	**87.7**	**71.7**	64.9	**66.2**	57.7	**64.2**	70.2 (+4.5)	25.3	43.2	79
A-PVT	89.6	68.1	32.6	61.3	60.0	77.9	**80.3**	90.8	65.5	87.4	69.5	62.3	64.6	58.1	54.9	68.2 (+2.5)	25.0	41.4	79
EIA-PVT	**90.0**	**77.5**	**43.8**	65.8	**62.6**	**81.2**	79.3	90.8	**78.2**	87.3	69.7	**78.0**	62.2	**70.6**	60.3	**73.2 (+7.5)**	25.5	30.8	68

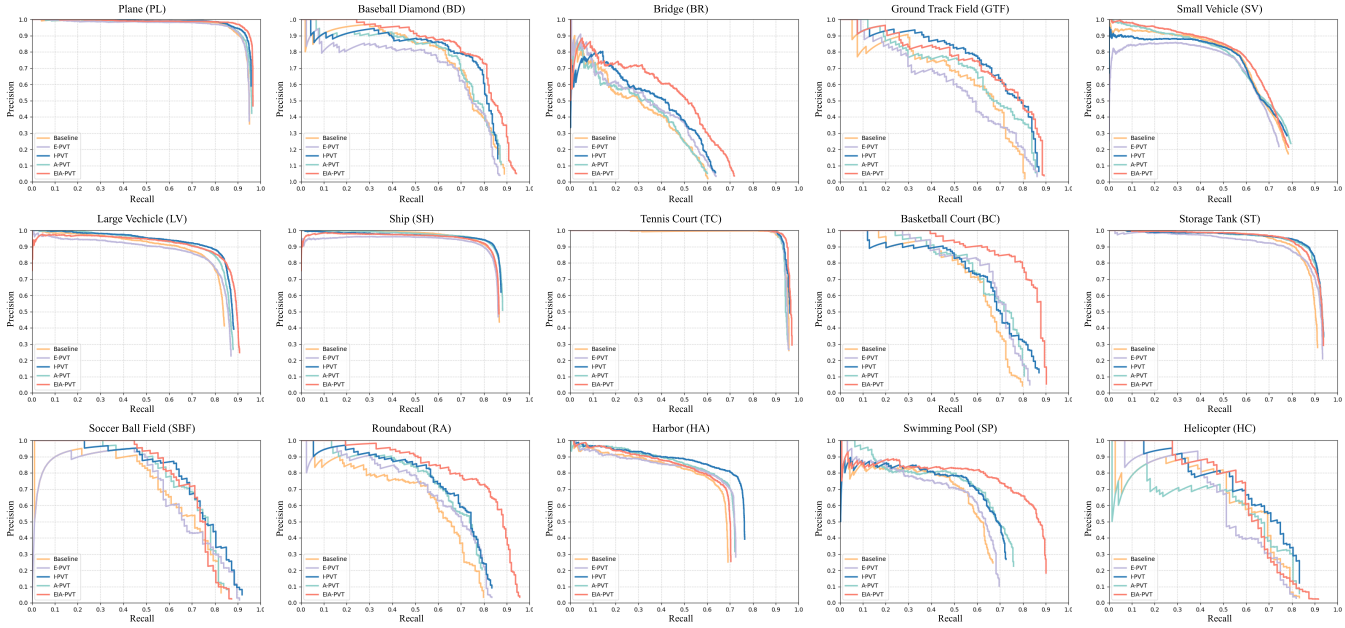


Fig. 8. PR curves of five different Transformer-based detection frameworks, i.e., the baseline, E-PVT, I-PVT, A-PVT, and EIA-PVT.

To provide detector variants for extensive comparison and proof of generalization, following PVT [23], we establish a series of EIA-PVT frameworks with different model scales, namely, EIA-PVT-Tiny, EIA-PVT-Small, EIA-PVT-Medium, and EIA-PVT-Large. The architectural details of these variants, mainly determined by the hyperparameters $\{N_i\}$, are listed as follows.

1) EIA-PVT-Tiny: $\{N_i\} = \{2, 2, 2, 2\}$.
2) EIA-PVT-Small: $\{N_i\} = \{3, 4, 6, 3\}$.
3) EIA-PVT-Medium: $\{N_i\} = \{3, 4, 18, 3\}$.
4) EIA-PVT-Large: $\{N_i\} = \{3, 8, 27, 3\}$.

And $\{C_i\} = \{64, 128, 320, 512\}$ is fixed for all variants, where $i \in \{1, 2, 3, 4\}$ means that there are four successive stages to form multiscale representations. In the experiments, unless otherwise specified, EIA-PVT-Small is employed as the default-oriented object detector and inherits some detection settings from PVT, such as the parameter initialization strategies and modified RetinaNet [14] as the detection head. All models are trained for 24 epochs on one NVIDIA GeForce RTX 3090 GPU, optimized by AdamW [86], with the initial learning rate of 1×10^{-3} and the weight decay of 1×10^{-3}, and the warm-up strategy is also implemented.

C. Ablation Studies

To verify the effectiveness and generalization of the proposed method, comprehensive ablation studies are conducted on the DOTA validation set in this section.

1) Effect of the Main Components of EIA-PVT: As described in Section III, unlike vision Transformers, EIA-PVT contains three main components, i.e., the **E**fficient adaptive multigrained routing mechanism, the **I**nductive bias encoding strategy, and the **A**ngle tokenization technique, each of which contributes to the entire framework, but can be implemented individually. We adopt the vanilla PVT to oriented object detection as the baseline method in the following experiments. To investigate the roles and effects of these three components, we add them separately to the vanilla PVT to form three reduced models, namely, E-PVT, I-PVT, and A-PVT, respectively. As shown in Table I, all reduced models achieve superior or comparable detection performance to the baseline. Particularly, thanks to the advantageous adaptive multigrained routing mechanism, our E-PVT not only maintains the detection accuracy but also dramatically improves the inference efficiency, with more than 10-G FLOPs reduction, accounting for 26.4% of the original computational load. Meanwhile, both I-PVT and A-PVT are able to boost the detection accuracy, bringing 4.5% and 2.5% mAP gains, respectively, at the expense of a small increase in Param and FLOPs,

which demonstrates their effectiveness in enhancing inductive bias and tokenizing orientation knowledge, respectively. The last row of Table I presents the performance of EIA-PVT, which contains all three components. It can be observed that EIA-PVT substantially outperforms the baseline method in all aspects, including detection accuracy, computational efficiency, and latency. Comparing the baseline model and E-PVT, they have almost the same model size in terms of Param, but their speeds in terms of FLOPs and latency are 41.3 G and 78 ms, respectively, for the baseline model, and 30.4 G and 67 ms, respectively, for E-PVT. This demonstrates the effectiveness of the proposed multigrained routing mechanism. For detection accuracy, EIA-PVT achieves 73.2% mAP, representing an improvement of 7.5% compared with the baseline model. It is worth noting that there is no linear correlation between the FLOPs of EIA-PVT and the reduced models, i.e., E-PVT, I-PVT, and A-PVT, since our proposed routing mechanism has dynamic adaptability and the token sparsity varies with different encoder architectures and training samples. In addition, we plot the PR curves of the above-mentioned five different Transformer models in Fig. 8, where across all 15 categories, our proposed frameworks consistently achieve the best detection performance. Among them, EIA-PVT outperforms all other models in most cases, while all the three reduced models, E-PVT, I-PVT, and A-PVT, demonstrate their respective superiority over the baseline. Based on the above-mentioned observations and analysis, the effectiveness and cooperation of the three main components have been demonstrated for remote sensing-oriented object detection.

Effect of Multiscale Representations: As illustrated in Fig. 2, there are four consecutive stages to generate hierarchical Transformer representations with different scales or resolutions, similar to the feature pyramid in CNNs. Table II validates the effect of multiscale representations on the baseline model and our EIA-PVT. It can be observed that with only single-scale features, the detection performance of both the baseline and EIA-PVT drops dramatically, achieving only 53.9% mAP and 60.2% mAP, respectively. It is worth noting that under the single-scale setting, keeping the spatial resolution of the features the same as the input image will lead to an extremely huge computational burden, exhausting all GPU memory. Consequently, to enable fair comparison of single-scale representations, the feature scale is uniformly shrunk to 1/16 of the original scale, resulting in comparable model parameters and computational complexity to the multiscale setting, as shown in Table II. The above-mentioned results have demonstrated that it is still necessary to construct multiscale representations in a hierarchical pyramid scheme for Transformers, which can benefit both detection performance and model efficiency.

2) Analysis of E-PVT: In this section, we further delve into E-PVT to interpret the effect of several key factors on the final detection performance and efficiency. The quantitative and qualitative results are reported in Tables III–V and Figs. 9–11.

Effect of the Granularity Candidate Sets in AMGR: To facilitate token sparsity and accelerate self-attention computation,

TABLE II

ABLATION EXPERIMENTS ON MULTISCALE REPRESENTATIONS IN THE BASELINE AND OUR EIA-PVT. "×" REPRESENTS THAT ONLY SINGLE-SCALE FEATURES ARE GENERATED THROUGHOUT THE WHOLE MODEL, WHILE "✓" DENOTES THE CONSTRUCTION OF MULTISCALE REPRESENTATIONS AS HIERARCHICAL OUTPUT

Methods	Multi-Scale Representations	mAP (Δ)	Param (Δ)	FLOPs (G)
Baseline	✓	65.7	**24.9M**	41.3
	×	53.9 (-11.8)	24.9M	42.8
EIA-PVT	✓	**73.2**	25.5M	**30.8**
	×	60.2 (-13.0)	25.5M	31.7

TABLE III

ABLATION STUDIES ON DIFFERENT CONFIGURATIONS OF THE GRANULARITY CANDIDATE SET IN AMGR

Methods	Adaptive Encoding	$\mathcal{G}$	mAP (Δ)	Param (Δ)	FLOPs (G)
Baseline	×	–	65.7	**24.9M**	41.3
E-PVT	✓	$\{0,1\}$	62.8 (-2.9)	24.9M (+8K)	30.3
		$\{1,2\}$	65.5 (-0.2)	24.9M (+8K)	31.1
		$\{1,2,4\}$	**65.8 (+0.1)**	24.9M (+13K)	30.4
		$\{1,2,4,5\}$	65.1 (-0.6)	24.9M (+17K)	29.6
		$\{1,2,4,5,10\}$	63.6 (-2.1)	24.9M (+21K)	**28.9**

E-PVT includes AMGR, followed by the vanilla Transformer encoders (rather than the proposed IDPM). Table III reports the detection results of E-PVT with different granularity candidate sets $\mathcal{G}$ in AMGR, which handle the granularity range and suggest the potential ceiling and floor of computational complexity. Without AMGR to diminish the spatial redundancy and perform adaptive encoding, i.e., all image regions indiscriminately yielding the finest grained patches and dense tokens, the baseline Transformer requires 41.3-G FLOPs. In contrast, with adaptive encoding, all E-PVT variants significantly reduce the computational cost. In the second row of Table III, when $\mathcal{G} = \{0, 1\}$, the granularity indices essentially degenerate into a learnable binary mask, where each pixel will be determined to remain unchanged or directly removed from the self-attention computation. However, this strategy was found to be suboptimal with serious performance degradation, with a 2.9% drop in mAP. It is noticed from the third row to the last row of Table III that by introducing coarse granularity candidates, E-PVT tends to derive coarser-grained patches via AMGR, resulting in a gradual decrease in computational cost. Particularly, when $\mathcal{G} = \{1, 2, 4\}$, the detection accuracy and efficiency can be ideally balanced, i.e., the computational complexity is significantly reduced without any deterioration in mAP. Therefore, this setting is utilized by default in other experiments. Additionally, as tabulated in the penultimate column of Table III, the newly introduced routing-related operations only bring negligible extra parameters, prevailing its portability.

Effect of the Computational Complexity Budget: To purposely constrain the expected computational complexity in the calculations of sparse token-based self-attention, a computational budget term is proposed and adopted in the overall training loss, as described in Section III-B. Table IV compares

TABLE IV

ABLATION EXPERIMENTS WITH DIFFERENT COMPUTATIONAL BUDGET CONSTRAINS σ IN E-PVT. HIGHER σ EMPIRICALLY INDICATES THAT HIGHER COMPUTATIONAL COMPLEXITY IS ACCEPTABLE

Computational Complexity Budget	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 0.75$	$\sigma = 1.0$
mAP (%)	55.9	63.1	**65.8**	65.7	**65.8**
Param (M)	24.9	24.9	24.9	24.9	24.9
FLOPs (G)	**21.8**	28.0	30.4	37.2	40.2

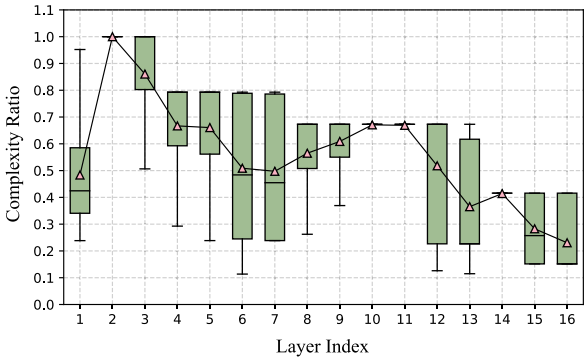


Fig. 9. Statistical distribution of computational complexity of all encoder layers in E-PVT-Small. After training, the 16 different encoder layers have their computational complexities fluctuated within different ranges, which further reflects that dynamic adaptive-grained encoding should be preferentially employed for each layer, while it is suboptimal to treat all layers equally by adopting static encoding.

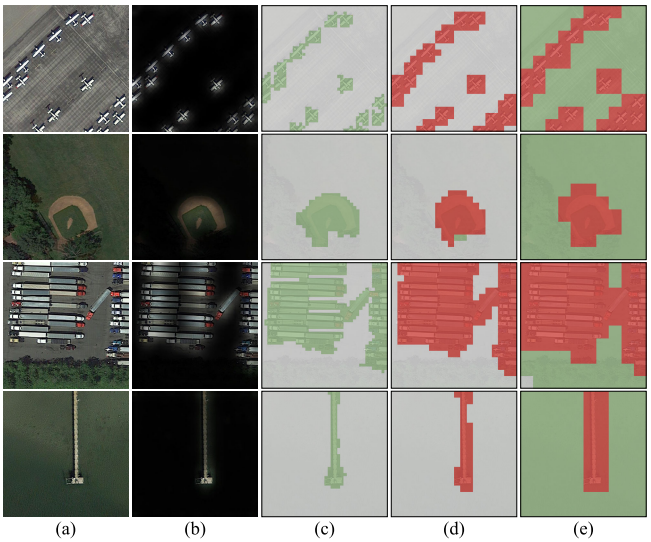


Fig. 10. Visualization of multigrained routing results. (a) Input images. (b) Attention mask. (c) Multigrained prediction map of stage 1 of E-PVT. (d) Multigrained prediction map of stage 2 of E-PVT. (e) Multigrained prediction map of stage 3 of E-PVT. The granularity candidate set is $\mathcal{G} = \{1, 2, 4\}$, whose three corresponding elements are highlighted in red, green, and gray, respectively. Larger (coarser) granularity typically indicates fewer tokens and lower computational complexity, and our method tends to allocate more tokens to discriminative foreground regions, thereby dramatically reducing the overall computational cost.

the final detection results and efficiencies by varying the budget constraint σ, where $\sigma \in \{0, 0.25, 0.5, 0.75, 1.0\}$. It can be seen that there is no noticeable mAP difference if σ is selected from $\{0.5, 0.75, 1.0\}$. However, fixing the budget to half of the original computation, i.e., $\sigma = 0.5$, leads to higher efficiency, which is taken as the default configuration for other experiments. Thanks to the proposed AMGR, even if $\sigma = 1.0$, the computational complexity is still slightly reduced to 40.2-G FLOPs, without loss of accuracy, compared with the original 41.3-G FLOPs. Actually, this phenomenon confirms that spatial redundancy in remote sensing scenarios allows our Transformers to yield comparable detection performance with much less computational cost. Nevertheless, the detection accuracy of E-PVT with σ fixed at 0 or 0.25 degrades, which may be caused by misassigning coarse-grained tokens to informative regions. In practice, it is unreasonable to set $\sigma = 0$, but here we attempt to explore the minimum requirement of the computational load, and the results imply that E-PVT achieves a detection performance of 55.9% mAP, with only about half of the previous FLOPs.

Effect of Layerwise Dynamic Adaptability: As interpreted earlier, E-PVT is equipped with powerful layerwise dynamic adaptability, and the computational complexity of its Transformer encoders varies with the input. Taking E-PVT-Small as an example, we provide the statistical distribution of the complexity ratios of the 16 encoder layers in Fig. 9. It can be observed that the computational complexity patterns of different encoder layers fluctuate considerably. This phenomenon verifies the necessity and superiority of dynamic token sparsity over static one since consecutive layers play distinct roles during encoding with varying information redundancies and correlations. From the trend of the complexity ratios in Fig. 9, we notice that deeper layers usually require less computation, in compliance with the hierarchical feature pyramid structure. Furthermore, despite the diversity of training samples, several encoder layers tend to consistently maintain higher computational complexity, which implies their unique indispensable functions in Transformer-based detectors.

Visualization of Adaptive Multigrained Routing: It has been emphasized that the proposed multigrained routing mechanism is adaptive to different samples, termed data dependency. Fig. 10 demonstrates this valuable property, where the multigrained routing processes are visualized with different images. Geospatial objects in remote sensing scenarios are typically characterized by various object textures and structures. For instance, the four categories of remote sensing objects in Fig. 10 have diverse scales, shapes, colors, orientations, and arrangements. To address this challenge, E-PVT is still able to adaptively nominate appropriate granularity candidates for different regions of each image. This is the key to balancing the detection performance and model efficiency. Specifically, as shown in Fig. 10(b), the self-attention mechanism essentially distinguishes informative foreground objects from the background, facilitating granularity indexing. As the number of encoders increases, the feature resolution becomes smaller, and the feature tokens change gradually from coarse-grained to fine-grained, as illustrated in Fig. 10(c)–(e). More importantly, compared with discriminative foreground regions, background regions are always assigned relatively coarser granularity, which allows generated sparse tokens for self-attention calculations. Furthermore, the predicted multigrained patterns are highly consistent with the corresponding attention masks,

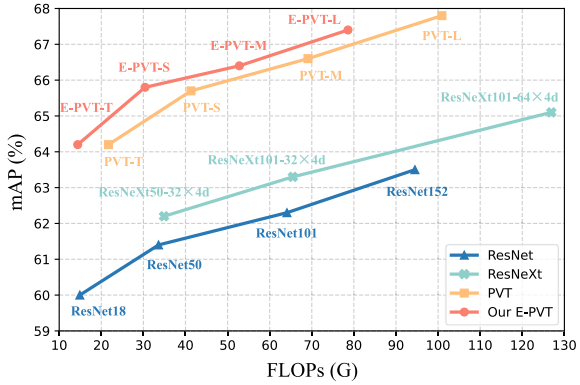


Fig. 11. Comparison of Transformer-based and CNN-based models in terms of detection accuracy and computational complexity. Our E-PVT and the vanilla PVT are Transformer-based models, while the ResNet [76] and ResNeXt [87] series serve as representatives of CNN models.

demonstrating the effective cooperation between our proposed routing mechanism and normal Transformer encoders.

Efficient Transformers versus CNNs: To verify the effectiveness and efficiency of our proposed method, Fig. 11 compares E-PVT with the vanilla PVT, ResNet [76], and ResNeXt [87], with different model sizes, where the latter two are commonly used famous CNN models. It can be observed that, PVT extensively exhibits better detection accuracy compared with ResNet, albeit at a higher computational cost, demonstrating that Transformers have a higher performance ceiling than CNNs in the task of remote sensing object detection. Therefore, our E-PVT inherits this ability, surpassing ResNet and ResNeXt in terms of accuracy and efficiency, even when being scaled up. To have comparable or even less computational cost, our E-PVT models outperform their CNN-based counterparts. More detailed experimental results are reported in Table V. For example, with similar computational complexity, E-PVT-Tiny achieves an mAP gain of 4.2% over ResNet-18, and its FLOPs are only 14.4 G, which is 33.6% lower than that of PVT-Tiny (21.7 G), yet without any accuracy drop. Similar patterns can be found for models of other sizes, where the high performance of Transformers is consistently maintained, while the computational cost is equal to or even lower than that of traditional CNNs. In addition, we further define a metric in Table V, termed (mAP/FLOPs), which represents the contribution of every GFLOP unit to the final mAP and is an important consideration in real-world applications. Thus, higher (mAP/FLOPs) should generally be preferred, given the computational cost constraint. For each of the four comparison groups in Table V, our E-PVT consistently achieves the best performance in terms of different evaluation criteria. These results validate the superiority of our proposed efficient Transformers models over CNN models and existing heavy Transformers.

3) Analysis of I-PVT: In this section, we analyze the inductive bias enhancement and dual-path encoding strategy based on I-PVT, and the relevant quantitative and qualitative results are illustrated in Table VI and Fig. 12, respectively.

Effect of Inductive Bias: As discussed in Section III, I-PVT identifies intrinsic inductive bias from two complementary

TABLE V

COMPARISON OF CNNS AND TRANSFORMERS WITH DIFFERENT MODEL SIZES. RESNET [76] AND RESNEXT [87] ARE ADOPTED AS THE CNN-BASED COUNTERPARTS. ALL METHODS ARE DIVIDED INTO FOUR GROUPS ACCORDING TO DIFFERENT MODEL SIZES. HIGHER (mAP/FLOPs) INDICATES HIGHER PRACTICAL VALUE FOR EACH GROUP

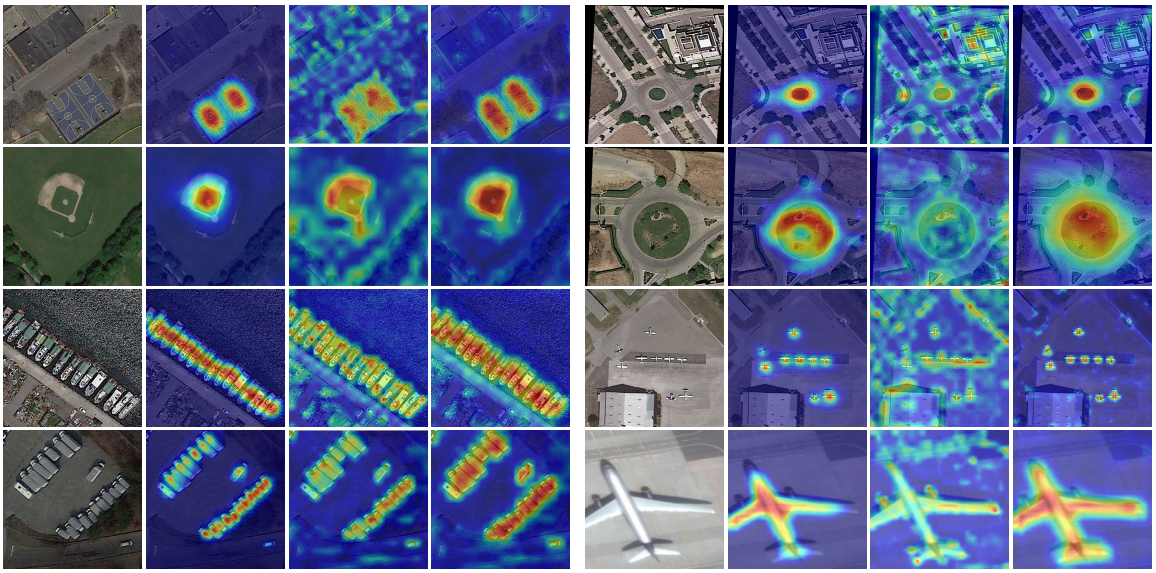
Models	Param (M)	FLOPs (G)	mAP (Δ)	$\frac{\text{mAP}}{\text{FLOPs}}\uparrow$
ResNet-18	**11.0**	14.9	60.0	4.03
PVT-Tiny	13.5	21.7	**64.2 (+4.2)**	2.96
E-PVT-Tiny (Ours)	13.5	**14.4**	**64.2 (+4.2)**	**4.46**
ResNet-50	23.3	33.6	61.4	1.83
ResNeXt-50-32×4d	**22.8**	34.9	62.2 (+0.8)	1.78
PVT-Small	24.9	41.3	65.7 (+4.3)	1.59
E-PVT-Small (Ours)	24.9	**30.4**	**65.8 (+4.4)**	**2.16**
ResNet-101	42.3	64.1	62.3	0.97
ResNeXt-101-32×4d	**41.9**	65.5	63.3 (+1.0)	0.97
PVT-Medium	44.7	69.1	**66.6 (+4.3)**	0.96
E-PVT-Medium (Ours)	44.8	**52.8**	66.4 (+4.1)	**1.26**
ResNet-152	**57.9**	94.5	63.5	0.67
ResNeXt-101-32×4d	81.0	126.9	65.1 (+1.6)	0.51
PVT-Large	62.0	100.9	**67.8 (+4.3)**	0.67
E-PVT-Large (Ours)	62.1	**78.6**	67.4 (+3.9)	**0.86**

TABLE VI

ABLATION STUDIES OF INDUCTIVE BIAS WITH DIFFERENT CONFIGURATIONS. § INDICATES REMOVING BATCH NORMALIZATION IN ILEP OF IDPM. † REPRESENTS USING THE SAME DILATION RATE SET IN FOUR SUCCESSIVE STAGES, WHILE ‡ DENOTES THAT THE PYRAMID SETTING IS EMPLOYED, I.E., $\mathcal{D}_1 = \{1, 2, 3, 4\}$, $\mathcal{D}_2 = \{1, 2, 3\}$, $\mathcal{D}_3 = \{1, 2\}$, AND $\mathcal{D}_4 = \{1\}$

Methods	Inductive Bias	Dilation $\mathcal{D}_i$	Dual-Path	DW-FFN	mAP (Δ)	Param (M)	FLOPs (G)
Baseline	×	–	–	–	65.7	**24.9**	**41.3**
I-PVT	✓	×	✓	×	68.1 (+2.4)	25.1	42.4
		×	✓§	×	67.7 (+2.0)	25.1	42.2
		×	✓	✓	68.8 (+3.1)	25.3	43.2
		$\{1, 2\}^{\dagger}$	×	×	66.4 (+0.7)	24.9	41.3
		$\{1, 2, 3\}^{\dagger}$	×	×	66.9 (+1.2)	24.9	41.3
		$\{1, 2, 3, 4\}^{\dagger}$	×	×	67.3 (+1.6)	24.9	41.3
		$\{1, 2, 3, 4, 5\}^{\dagger}$	×	×	66.7 (+1.0)	24.9	41.3
		$\{1, 2, 3, 4\}^{\ddagger}$	×	×	67.8 (+2.1)	24.9	41.3
		$\{1, 2, 3, 4\}^{\ddagger}$	✓	×	69.3 (+3.6)	25.1	42.4
		$\{1, 2, 3, 4\}^{\ddagger}$	✓	✓	**70.2 (+4.5)**	25.3	43.2

perspectives, i.e., the scale invariance in space brought by PPE and the locality introduced by IDPM. Specifically, in PPE, the scale-invariance inductive bias is mainly provided by the proposed pyramid projection based on multidilated convolutions, while in IDPM, the locality inductive bias is first enhanced by dual-path modulation and then by DW-FFN. To figure out the impact of these factors on the final detection performance, we tabulate the detailed ablation results in Table VI. It is worth noting that any increase in inductive bias can benefit the detection accuracy, from a small mAP gain of 0.7% to a larger mAP gain of 4.5%. Thus, all I-PVT variants surpass the baseline method without enhancing inductive bias. The second and third rows investigate the separate effect of the dual-path encoding strategy, based on different configurations, i.e., with or without batch normalization in ILEP in Fig. 6. Then, 2.4% and 2.0% mAP improvements can be observed, verifying the effectiveness of modeling local information and global dependencies in parallel and the necessity of batch normalization. If further equipped with DW-FFN, I-PVT



(a) (b)

Fig. 12. Visualization of attention flow in I-PVT. (a) Visual comparison based on different remote sensing object categories; from top to bottom are BCs, BDs, SHs, and SVs. (b) Visual comparison based on the same object categories yet at different scales, from top to bottom are smaller RAs, larger roundabouts, smaller PLs, and larger PLs. For (a) and (b), from left to right are the input images, class activation maps of ILEP in I-PVT, the attention maps of the original tokens in the vanilla Transformer encoder of the baseline, and the attention maps of the inductive tokens processed by IDPM in I-PVT.

achieves better performance, reaching 68.8% mAP. Both the dual-path strategy and DW-FFN aim to enhance locality, which demonstrates the importance of this form of inductive bias for remote sensing object detection. Next, from the fifth to the ninth rows, we explore the effect of dilated convolutions with different settings. Compared with the baseline model, the scale-invariance inductive bias based on the dilated convolutions consistently improves the detection accuracy, and the best two results are achieved when $\mathcal{D}_i = \{1, 2, 3, 4\}$. More importantly, if the dilation rate set follows a pyramid pattern, i.e., its cardinality decreases by stages, a detection accuracy up to 67.8% can be observed in the last third row of Table VI. Naturally, based on this configuration, by gradually including dual-path modulation and DW-FFN, I-PVT can deliver more performance gains, ultimately reaching an mAP of 70.2%, the highest among all results. It is noteworthy that only a few additional parameters and FLOPs increments are introduced in I-PVT, which is acceptable in terms of the remarkable accuracy improvements. The above-mentioned results demonstrate the effectiveness and superiority of our strategies in enhancing inductive bias for remote sensing object detection.

Visualization of Attention Flow in I-PVT: To further investigate the properties and benefits of inductive bias included in I-PVT, we provide some qualitative visualization results in Fig. 12. Grad-CAM [88] and the attention rollout method [89] are applied to visualize the class activation maps for convolutions and the attention maps for Transformer tokens, respectively. It can be observed that our I-PVT generally has a stronger ability to distinguish and locate target objects under complex background interference than the baseline model. Specifically, the attention maps of the vanilla Transformer encoders [the third columns of Fig. 12(a) and (b)] accumulate considerable long-distance relational redundancy, essentially caused by the purposeless computation of self-attention. They may appear as global noise and impair detection performance. In contrast, the attention maps computed based on the enhanced inductive tokens in I-PVT [the fourth columns of Fig. 12(a) and (b)] show more robust feature dependencies, where the attentive regions are more precise and complete, while the uninformative background is significantly suppressed. This improvement is mainly attributed to the introduction of inductive bias. As illustrated in the class activation maps [the second columns of Fig. 12(a) and (b)], the ILEP of IDPM tends to activate only narrow or even incomplete local regions and can adopt this intrinsic locality to modulate the attention distribution in Transformers, resulting in more discriminative tokens or features for object detection. Moreover, large variance in object scales is considered to be a common challenge in remote sensing scenarios. As depicted in Fig. 12(b), our method can accurately adapt to various scales, which also validates the effectiveness of scale invariance as inductive bias. The above-mentioned observations and analysis demonstrate that the visual grounding capability of our I-PVT has been significantly enhanced by introducing appropriate inductive bias.

TABLE VII
ABLATION STUDIES ON ANGLE TOKENIZATION IN THE PROPOSED ATM

Methods	Decoupled	Stage 1	Stage 2	Stage 3	Stage 4	mAP (Δ)	Param (M)	FLOPs (G)
Baseline	×	–	–	–	–	65.7	**24.9**	**41.3**
A-PVT	✓	✓	×	×	×	66.7 (+1.0)	24.9	41.3
	✓	✓	✓	×	×	67.2 (+1.5)	24.9	41.4
	✓	✓	✓	✓	×	67.9 (+2.2)	25.0	41.4
	✓	✓	✓	✓	✓	**68.2 (+2.5)**	25.0	41.4

4) Analysis of A-PVT: The major difference between A-PVT and PVT lies in that ATM is introduced in each stage, and minor modifications are also made in the detection head to

TABLE VIII
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DIOR-R DATASET

Methods	APL	APO	BF	BC	BR	CH	DAM	ETS	ESA	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP	Param (M)	FLOPs (G)
RetinaNet-O [14]	61.5	28.5	73.6	81.2	24.0	72.5	19.9	72.4	58.2	69.3	79.5	32.1	44.9	77.7	67.6	61.1	81.5	47.3	38.0	60.2	57.6	36.5	133.4
Faster R-CNN-O [11]	62.8	26.8	71.7	80.9	34.2	72.6	19.0	66.5	65.8	66.6	79.2	35.0	48.8	81.1	64.3	71.2	81.4	47.3	50.5	65.2	59.5	41.1	134.4
Gliding Vertex [92]	65.4	28.9	75.0	81.3	33.9	74.3	19.6	70.7	64.7	72.3	78.7	37.2	49.6	80.2	69.3	61.1	81.5	44.8	47.7	65.0	60.1	41.3	134.6
RoI Transformer [4]	63.3	37.9	71.8	87.5	40.7	72.6	26.9	78.7	68.1	69.0	82.7	47.7	55.6	81.2	78.2	70.3	81.6	54.9	43.3	65.5	63.9	55.2	148.4
QPDet [93]	63.2	41.4	72.0	**88.6**	41.2	72.6	28.8	**78.9**	69.0	70.1	83.0	47.8	55.5	81.2	72.2	62.7	**89.1**	58.1	43.4	65.4	64.2	41.1	134.4
AOPG [50]	62.4	37.8	71.6	87.6	40.9	72.5	31.1	65.4	**78.0**	73.2	81.9	42.3	54.5	81.2	72.7	71.3	81.5	60.0	52.4	**70.0**	64.4	41.7	134.4
FASSROD [94]	61.5	36.3	72.2	87.9	42.1	71.0	31.5	70.2	75.7	**74.2**	81.8	45.6	55.0	82.0	73.4	70.9	81.5	**61.0**	51.0	68.9	64.8	41.9	175.6
CGCDet [91]	68.5	38.3	79.1	86.2	39.0	73.5	26.8	66.0	74.7	67.5	**84.5**	**48.0**	56.1	81.3	79.3	72.2	88.6	50.7	51.7	65.8	64.9	55.2	148.4
DODet [68]	63.4	**43.4**	72.1	81.3	**43.1**	72.6	**33.3**	78.8	70.8	**74.2**	75.5	**48.0**	**59.3**	**85.4**	74.0	71.6	81.5	55.5	51.9	66.4	65.1	69.4	397.7
EIA-PVT-Small (Ours)	**80.4**	36.5	**80.3**	81.3	40.5	**80.5**	25.1	77.8	65.4	72.1	83.7	35.9	55.0	81.0	**80.7**	**77.9**	88.2	53.1	**54.0**	65.6	**65.8**	**27.6**	**84.3**

Fig. 13. Some qualitative detection results of the proposed EIA-PVT on the DIOR-R dataset.

TABLE IX
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DOTA DATASET

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
FR-O [47]	79.4	77.1	17.7	64.1	35.3	38.0	37.2	89.4	69.6	59.3	50.3	52.9	47.9	47.4	46.3	54.1
RoI Transformer [4]	88.6	78.5	43.4	75.9	68.8	73.7	83.6	90.7	77.3	81.5	58.4	53.5	62.8	58.9	47.7	69.6
CAD-Net [95]	87.8	82.4	49.4	73.5	71.1	63.5	76.6	90.9	79.2	73.3	48.4	60.9	62.0	67.0	62.2	69.9
O^2DNet [96]	89.3	82.1	47.3	61.2	71.3	74.0	78.6	90.8	82.2	81.4	60.9	60.2	58.2	67.0	61.0	71.0
SCRDet [97]	90.0	80.7	52.1	68.4	68.4	60.3	72.4	90.9	87.9	86.9	65.0	66.7	66.3	68.2	65.2	72.6
R^3Det [53]	89.2	80.8	51.1	65.6	70.7	76.0	78.3	90.8	84.9	84.4	65.1	57.2	68.1	69.0	60.9	72.8
CFC-Net [85]	89.1	80.4	52.4	70.0	76.3	78.1	87.2	90.9	84.5	85.6	60.5	61.5	67.8	68.0	50.1	73.5
APE [98]	89.7	76.8	51.3	71.7	73.1	77.2	79.5	90.8	79.0	84.5	66.5	64.7	74.0	67.7	58.4	73.7
ROSOOD [99]	88.7	79.4	52.3	65.5	74.7	80.8	87.4	90.8	84.3	83.4	62.6	58.1	67.0	72.3	69.3	74.4
AOPG [50]	89.3	83.5	52.5	70.0	73.5	82.3	88.0	90.9	87.6	84.7	60.0	66.1	74.2	68.3	57.8	75.2
DODet [68]	89.3	84.3	51.4	71.0	79.0	82.9	88.2	90.9	86.9	84.9	62.7	67.6	75.5	72.2	45.5	75.5
CBDA-Net [100]	89.2	85.9	50.3	65.0	77.7	82.3	87.9	90.5	86.5	85.9	66.9	66.5	67.4	71.3	62.9	75.7
QPDet [93]	89.6	83.7	54.1	73.9	78.9	83.1	88.3	90.9	86.6	84.8	62.0	65.6	74.2	70.1	58.2	76.3
CoF-Net [2]	89.6	83.1	48.3	73.6	78.2	83.0	86.7	90.2	82.3	86.6	67.6	64.6	74.7	71.3	78.4	77.2
CGCDet [91]	88.9	84.5	53.9	78.6	78.5	82.5	87.9	90.9	87.5	84.8	65.6	63.5	76.2	71.6	65.3	77.3
GF-CSL [101]	89.8	86.2	49.9	73.0	78.6	81.8	88.0	90.8	87.9	86.2	63.0	66.2	77.1	79.3	65.3	77.5
EIA-PVT-Small (Ours)	89.9	86.4	52.4	74.9	77.9	83.6	87.9	90.9	86.2	84.3	68.1	66.2	76.5	78.3	75.0	**78.6**

facilitate angle learning of ATM. Hence, we mainly evaluate the effect of the proposed angle tokenization technique.

Effect of Angle Tokenization: Table VII tabulates the ablation results for different insertion positions of ATM in A-PVT. Compared with the baseline model, all A-PVT variants exploit decoupled angle heads to enable explicit orientation knowledge for ATMs, without supervising inference from classification and localization. Despite information aggregations in the decoupled angle head, ATMs in consecutive stages are essentially separate, which allows them to be inserted into arbitrary stages according to preference. It can be noticed from the last row of Table VII that introducing an ATM to each stage of A-PVT derives the best detection results, at 68.2% mAP, which is 2.5% higher than the baseline model. We infer that since the tokens or features at different stages are responsible for detecting objects of different scales, it is reasonable and necessary to insert an ATM into each stage, which iteratively embeds orientation knowledge into the tokens and promotes angle representations for remote sensing object detection. In addition, thanks to the use of

Fig. 14. Some qualitative detection results of the proposed EIA-PVT on the DOTA dataset.

lightweight group convolutions [90] in ATMs, A-PVT only introduces negligible extra model parameters and computational complexity, without obvious efficiency sacrifice. The above-mentioned results demonstrate the effectiveness and superiority of our proposed angle tokenization technique for remote sensing-oriented object detection.

D. Comparative Experiments With State-of-the-Art Methods

In this section, we compare the proposed EIA-PVT with other state-of-the-art methods on two popular remote sensing-oriented object detection datasets: DIOR-R and DOTA.

1) Results on DIOR-R: We evaluate and compare the proposed framework with other state-of-the-art methods on the largest DIOR-R dataset, and the results are listed in Table VIII. It can be observed that our EIA-PVT-Small outperforms all other CNN-based detectors in terms of accuracy, which verifies the advancement and superiority of vision Transformers over CNNs on this task. Specifically, our method achieves 65.8% mAP across all categories, surpassing two very recent concurrent approaches, CGCDet [91] and DODet [68]. EIA-PVT-Small achieves the best detection performance in six categories and is also competitive in the remaining categories. The above-mentioned comparison further demonstrates the effectiveness of our method in facilitating Transformers for oriented object detection in remote sensing images. Some visual detection results are shown in Fig. 13.

2) Results on DOTA: As tabulated in Table IX, performance comparisons are also conducted on the DOTA dataset with 16 representative CNN-based methods. Similar to DIOR-R, the proposed EIA-PVT-Small framework outperforms all other counterparts on the DOTA dataset, including the very recent methods CoF-Net [2], QPDet [93], CGCDet [91], and GF-CSL [101]. Moreover, across all 15 categories, our Transformer-based detector consistently obtains the best or most favorable results, beating most existing detectors, demonstrating the generalization and superiority of our method for remote sensing object detection. Some qualitative detection results of the proposed EIA-PVT are presented in Fig. 14.

V. Conclusion

This article proposes a novel pyramid Transformer framework, named EIA-PVT, for oriented object detection in remote sensing imagery, which achieves high efficiency and high accuracy. Previous methods have failed to fully explore the utility and superiority of vision Transformers over CNNs in the task of remote sensing object detection, mainly due to three deficiencies, i.e., high computational cost, lack of inductive bias, and arbitrary orientation. To pave the way, EIA-PVT presents an efficient inductive Transformer framework with angle tokenization, constituted of four successive stages to form a pyramid hierarchy. Each stage contains a PPE, several EIEs, and an ATM. The core component EIE aims to reduce computational complexity and improve encoding efficiency by facilitating token sparsity and inducing global and local semantic relations by enhancing inductive bias, while ATM explicitly discriminates and embeds direction knowledge of oriented objects in remote sensing scenarios into Transformer tokens to promote angle learning. Extensive experiments were conducted on two public datasets, DOTA and DIOR-R, and the results have verified the effectiveness and efficiency of our proposed method, as well as its superiority over other state-of-the-art detectors.

References

[1] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.

[2] C. Zhang, K.-M. Lam, and Q. Wang, "CoF-Net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600617.

[3] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag. Replaces Newsletter*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[4] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2849–2858.

[5] T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, and L. Li, "Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608720.

[6] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624915.

[7] Q. Wang, X. Lu, C. Zhang, Y. Yuan, and X. Li, "LSV-LP: Large-scale video-based license plate detection and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 752–767, Jan. 2023.

[8] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag. Replaces Newsletter*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[9] 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.*, Jun. 2014, pp. 580–587.

[10] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[11] 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.

[12] 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. (CVPR)*, Jun. 2016, pp. 779–788.

[13] 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. (CVPR)*, Jul. 2017, pp. 936–944.

[14] 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. (ICCV)*, Oct. 2017, pp. 2999–3007.

[15] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[16] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[17] W. Zhang, L. Jiao, Y. Li, Z. Huang, and H. Wang, "Laplacian feature pyramid network for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604114.

[18] X. Lu, Y. Zhang, Y. Yuan, and Y. Feng, "Gated and axis-concentrated localization network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 179–192, Jan. 2020.

[19] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "ASSD: Feature aligned single-shot detection for multiscale objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607117.

[20] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[21] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[22] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[23] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[24] Y. Yuan et al., "HRFormer: High-resolution vision transformer for dense predict," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 7281–7293.

[25] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "ViTAE: Vision transformer advanced by exploring intrinsic inductive bias," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 28522–28535.

[26] H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, and P. Molchanov, "A-ViT: Adaptive tokens for efficient vision transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10799–10808.

[27] L. Song et al., "Dynamic grained encoder for vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 5770–5783.

[28] K. Li et al., "Uniformer: Unified transformer for efficient spatial–temporal representation learning," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–19.

[29] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. NeurIPS*, 2021, pp. 1–12.

[30] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2286–2296.

[31] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[32] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 357–366.

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[34] 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.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[35] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–26.

[36] Z. Chen, J. Zhang, and D. Tao, "Recurrent glimpse-based decoder for detection with transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5250–5259.

[37] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Hyperspectral remote sensing image subpixel target detection based on supervised metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4955–4965, Aug. 2014.

[38] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[39] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[40] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609413.

[41] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8122–8131.

[42] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "TCTrack: Temporal contexts for aerial tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14778–14788.

[43] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.

[44] Z. Xu et al., "RNGDet: Road network graph detection by transformer in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707612.

[45] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.

[46] Y. Zheng, P. Sun, Z. Zhou, W. Xu, and Q. Ren, "ADT-Det: Adaptive dynamic refined single-stage transformer detector for arbitrary-oriented object detection in satellite optical imagery," *Remote Sens.*, vol. 13, no. 13, p. 2623, Jul. 2021.

[47] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[48] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.

[49] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.

[50] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625411.

[51] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3500–3509.

[52] T. Zhang et al., "Foreground refinement network for rotated object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610013.

[53] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, and T. He, "SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2384–2399, Feb. 2023.

[54] X. Yang and J. Yan, "On the arbitrary-oriented object detection: Classification based approaches revisited," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1340–1365, May 2022.

[55] X. Yang et al., "Detecting rotated objects as Gaussian distributions and its 3-D generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4335–4354, Apr. 2023.

[56] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[57] J. Han et al., "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogramm. Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.

[58] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 673–685, Feb. 2021.

[59] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.

[60] H. He, Y. Lin, F. Chen, H.-M. Tai, and Z. Yin, "Inshore ship detection in remote sensing images via weighted pose voting," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3091–3107, Jun. 2017.

[61] H. Zhou, L. Wei, C. P. Lim, D. creighton, and S. Nahavandi, "Robust vehicle detection in aerial images using bag-of-words and orientation aware scanning," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7074–7085, Dec. 2018.

[62] J. Li, H. Zhang, R. Song, W. Xie, Y. Li, and Q. Du, "Structure-guided feature transform hybrid residual network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610713.

[63] 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, 2022, Art. no. 5614914.

[64] 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.

[65] G. Wang et al., "FSoD-Net: Full-scale object detection from optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602918.

[66] Q. Lin, J. Zhao, B. Du, G. Fu, and Z. Yuan, "MEDNet: Multiexpert detection network with unsupervised clustering of training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4703114.

[67] J. Wang, W. Yang, H.-C. Li, H. Zhang, and G.-S. Xia, "Learning center probability map for detecting objects in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4307–4323, May 2021.

[68] G. Cheng et al., "Dual-aligned oriented detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618111.

[69] X. Yang, Q. Liu, J. Yan, A. Li, Z. Zhang, and G. Yu, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3163–3171.

[70] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A general Gaussian heatmap label assignment for arbitrary-oriented object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 1895–1910, 2022.

[71] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2019, pp. 4171–4186.

[72] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[73] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[74] M. Liu, Q. Shi, J. Li, and Z. Chai, "Learning token-aligned representations with multimodel transformers for different-resolution change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4413013.

[75] L. Chen, R. Luo, J. Xing, Z. Li, Z. Yuan, and X. Cai, "Geospatial transformer is what you need for aircraft detection in SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5225715.

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

[77] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[78] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and X. Li, "GCWNet: A global context-weaving network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619912.

[79] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[80] A. Veit and S. Belongie, "Convolutional networks with adaptive inference graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.

[81] T. Verelst and T. Tuytelaars, "Dynamic convolutions: Exploiting spatial sparsity for faster inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2317–2326.

[82] Z. Xie, Z. Zhang, X. Zhu, G. Huang, and S. Lin, "Spatially adaptive inference with stochastic feature sampling and interpolation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 531–548.

[83] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–12.

[84] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[85] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605814.

[86] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–12.

[87] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[88] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[89] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4190–4197.

[90] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving CNN efficiency with hierarchical filter groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5977–5986.

[91] Y. Wang et al., "Learning oriented object detection via naive geometric computing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 10, 2023, doi: 10.1109/TNNLS.2023.3242323.

[92] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[93] Y. Yao et al., "On improving bounding box representations for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600111.

[94] Y. Yuan, Z. Li, and D. Ma, "Feature-aligned single-stage rotation object detection with continuous boundary," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5538011.

[95] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[96] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.

[97] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8231–8240.

[98] Y. Zhu, J. Du, and X. Wu, "Adaptive period embedding for representing oriented objects in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7247–7257, Oct. 2020.

[99] L. Hou, K. Lu, and J. Xue, "Refined one-stage oriented object detection method for remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 1545–1558, 2022.
[100] S. Liu, L. Zhang, H. Lu, and Y. He, "Center-boundary dual attention for oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603914.
[101] J. Wang, F. Li, and H. Bi, "Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707013.

Cong Zhang (Graduate Student Member, IEEE) received the B.E. degree from the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, in 2018, and the M.E. degree from the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, in 2021. He is currently pursuing the Ph.D. degree with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong.

His research interests include remote sensing and computer vision.

Jingran Su received the B.E. degree from the School of Automation, Northwestern Polytechnical University, Xi'an, China, in 2018, and the M.E. degree from the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, in 2021. He is currently pursuing the Ph.D. degree in computer science with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong.

His research interests include computer vision and machine learning.

Yakun Ju (Member, IEEE) received the B.Sc. degree from the School of Mechanical Engineering, Sichuan University, Chengdu, China, in 2016, and the Ph.D. degree in computer science and technology from the Ocean University of China, Qingdao, China, in 2022, under the supervision of Prof. Junyu Dong.

He was a Research Assistant with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, in 2021, where he is currently a Post-Doctoral Fellow, working with Prof. Kenneth K. M. Lam. His research interests include computer vision, deep learning, and image processing.

Kin-Man Lam (Senior Member, IEEE) received his Associateship (Hons.) in electronic engineering from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic), Hong Kong, in 1986, the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering, The University of Sydney, Camperdown, NSW, Australia, in 1996.

From 1990 to 1993, he was a Lecturer with the Department of Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, as an Assistant Professor in 1996, where he became an Associate Professor in 1999 and has been a Professor since 2010. He is also an Associate Dean of the Faculty of Engineering, The Hong Kong Polytechnic University. He was actively involved in professional activities. His research interests include image processing, computer vision, and human face analysis and recognition.

Prof. Lam was the Director of the Student Services and the Director of the Membership Services of the IEEE Signal Processing Society from 2012 to 2014 and from 2015 to 2017, respectively. He was also the Vice President of the Member Relations and Development and the Publications of the Asia–Pacific Signal and Information Processing Association (APSIPA) from 2014 to 2017 and from 2017 to 2021, respectively. He has been a member of the organizing committee or the program committee of many international conferences. He is currently the Member-at-Large of APSIPA. He serves as a Senior Editorial Board Member for the *APSIPA Transactions on Signal and Information Processing*. He was the Chairperson of the IEEE Hong Kong Chapter of Signal Processing from 2006 to 2008. He was the General Co-Chair of the 2012 IEEE International Conference on Signal Processing, Communications and Computing, Hong Kong, the APSIPA Annual Summit 2015, Hong Kong, and the 2017 IEEE International Conference on Multimedia and Expo, Hong Kong, and the Technical Chair of the 2020 IEEE International Conference on Visual Communications and Image Processing. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2009 to 2014 and *Digital Signal Processing* from 2014 to 2018. He was also an Editor of the *HKIE Transactions* from 2013 to 2018 and an Area Editor of *IEEE Signal Processing Magazine* from 2015 to 2017. He serves as an Associate Editor for the *EURASIP International Journal on Image and Video Processing*.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi'an, China, where he is also with the Key Laboratory of Intelligent Interaction and Applications, Ministry of Industry and Information Technology. His research interests include computer vision and pattern recognition.