EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

FrequencyFormer: Oriented Object Detection with Frequency Transformer

Shuai Liu[1,2], Haiming Wang[1], Zhibin Li[3,4,*],Peiyang Wei[5]

[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin, China
[2]Tianjin Institute of Advanced Technology, Tianjin, China
[3]Chengdu University of Information Technology, Chengdu, China
[4]Sichuan University of Arts and Science, Dazhou, China
[5]Chengdu University of Information Technology, Chengdu, China

Abstract

Detecting objects with oriented bounding boxes have shown impressive generalizations in the challenging scenes with densely packed objects with arbitrary rotations. Existing oriented object detectors rely on customized operations like anchor pre-definition and NMS post-processing for accuracy improvement. However, those components usually bring extensive computational costs and complicate the pipeline, and thus limit the scalability of existing methods. In this paper, we propose a new paradigm, FrequencyFormer, for end-to-end oriented object detection. Upon the Transformer based encoder-decoder framework, two key ingredients are proposed to adapt it to detect oriented objects robustly. First, a frequency boosted query update strategy is designed to enhance the shape encoding of object queries by incorporating the frequency vectors of oriented objects. Second, a dynamic matching strategy is introduced to facilitate the training process, in which the matching weights are adjusted adaptively as the training progress. Experimental results on DOTA and HRSC2016 datasets demonstrate that our FrequencyFormer achieves competitive performance with state-of-the-art methods.

Received on 23 October 2025; accepted on 28 November 2025; published on 02 December 2025

Keywords: Transformer, Oriented object, Frequency

doi:10.4108/airo.10701

1. Introduction

Oriented object detection aims to localize objects with enclosed, oriented bounding boxes. Compared with the general object detection in horizontal perceptive, oriented bounding boxes are effective at discriminating densely packed objects with arbitrary directions, thus showing impressive performance in object detection remote sensing object detection [1] and scene text detection [2].

Thanks to the recent advances [3, 4] in general object detection, various methods [5–9] have been developed for oriented object detection and achieved promising results. Some early works [5, 6] follow the two-stage paradigm [3, 4] by introducing one additional angle dimension to the original horizontal anchors. However, the two-stage methods always define hundreds of rotated anchors in various shapes, sizes and angles to alleviate the sensitivity for object detection, which incur undesirable computational costs. To tackle this issue, some attempts [7–9] have been made to develop a one-stage framework by getting rid of the anchor computation process, namely anchor-free methods. The anchor-free detectors show impressive performance on runtime speedup, and have become the mainstream in oriented object detection. However, to distill the matched boxes from noisy predictions, the anchor-free methods usually need to design a series of complex post-processing like Non-Maximum Suppression (NMS) for rotated boxes.

To simplify the detection pipeline, a recent work [10] proposes to tackle object detection as a set prediction problem and constructs an end-to-end framework called DETR. The DETR uses Transformer [11] to

*Corresponding author, Email: Lizhibin111@outlook.com.

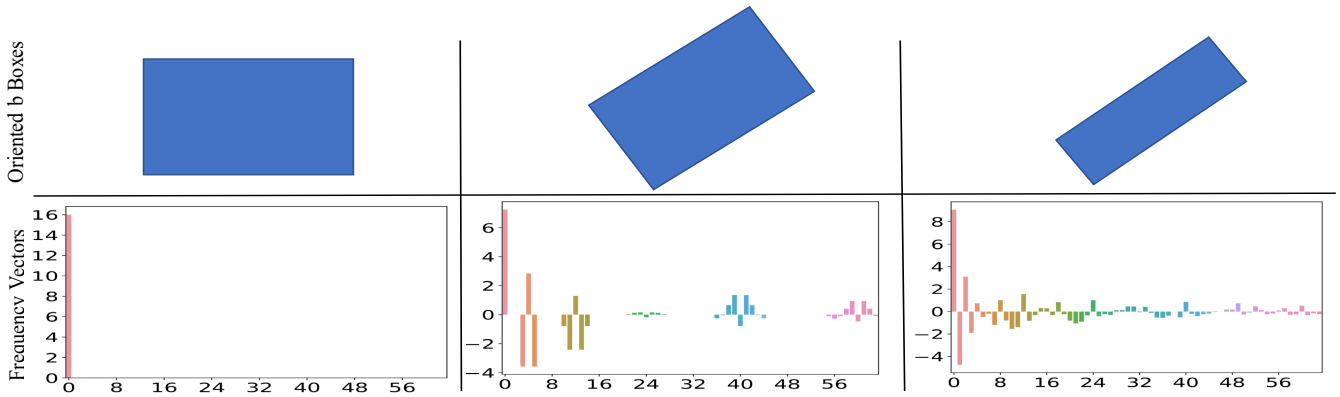

Figure 1. Frequency of oriented objects with different orientations and aspect ratios. Top: oriented bounding boxes in different rotated angles aspect ratios. Bottom: the corresponding frequency vectors.

strength the interactions between object queries and image content. With the success of DETR, a number of follow-ups [12, 13] are proposed for further improvement on detection accuracy or adaptation to other tasks. It is then natural to ask whether we can leverage the merit of Transformer to build an end-to-end paradigm for oriented object detection? One simple solution is to extend the original DETR [10] by adding one extra dimension to the box regression branch. However, we found that such straightforward extension does not adapt well in oriented object detection. DETR incorporates the visual features with positional embedding to enhance the spatial structure in the attention formulation of Transformer. The positional encoding shows great importance in recognizing objects by horizontal bounding boxes. However, due to the lack of shape prior (*e.g.*, size, aspect ratio and rotation angle) for rotated objects, the simple incorporation of positional embedding may not generalize well for the regression of oriented boxes.

Compared with the positional embedding in spatial domain, we find that the representations of oriented objects under different shapes show discrimintative variations in frequency domain. As we can see in Fig. 1, the frequency vectors of the same bounding box in various rotated angles and aspect ratios show large difference. The frequency representation is capable of compressing the 2-D oriented boxes into 1-D vectors, which can be incorporated into Transformer to construct robust spatial embedding for accurate oriented object detection.

With the above observation, we propose a new paradigm, FrequencyFormer, for oriented object detection in end-to-end manner. We introduce the frequency representations into Transformer to boost the localization of oriented objects in arbitrary direction. The FrequencyFormer is built upon the Transformer-based encoder-decoder architecture [10], in which a fixed set of learned queries are interacted with image content for object classification and localization. A frequency boosted query update strategy is designed to enhance the shape encoding of object queries by incorporating the frequency vectors. Specifically, in each decoder layer of Transformer, we first map the predicted oriented objects into frequency domain via discrete cosine transform. The obtained frequency vectors, which contain discriminative information for distinguishing various oriented objects, will be aggregated with object queries to boost the object detection in next decoder layer. With this query update strategy, the effective shape encoding are incorporated with image content in a progressive manner for robust oriented object detection. Besides, to facilitate the model training, a dynamic matching strategy is proposed by extending the bipartite matching loss with adaptive weights adjusted during training progress.

We evaluate the proposed FrequencyFormer on two public benchmarks including DOTA [14] and HRSC2016 [15]. The experimental results demonstrate that our method achieves competitive performance against state-of-the-art approaches, without applying any costumed post-processing.

To summarize, we propose FrequencyFormer, a fully end-to-end paradigm for oriented object detection. This is enabled by two key components, namely the frequency boosted query update strategy to incorporate object query with robust shape encoding from frequency domain, and the dynamic matching strategy to facilitate the end-to-end training of network. It results in a simplified pipeline for directly outputting predictions without any customized post-processing. The experimental results on two datasets demonstrate that the proposed FrequencyFormer performs favorably against state-of-the-art methods.

2. Related Work

2.1. Oriented Object Detection

Oriented object detection is widely used in remote sensing object detection [7, 9] and scene text detection [16], in which objects are often densely packed with arbitrary directions. Many early methods leverage the advances of general object detection by introducing one additional angle dimension to the horizontal bounding boxes. Generally, existing methods in oriented object detection can be classified into two categories, called anchor-based and anchor-free methods.

In anchor-based methods, a series of rotation anchors with arbitrary orientations and aspect ratios are designed to provide prior information for oriented objects. For example, Ding *et al.* [6] propose a rotated RoI transformer module to solve the misalignment between oriented objects and image features. In anchor-free methods, the complex anchor design is removed and the oriented bounding boxes are obtained by box regression on pixel wise directly. For example, Liu *et al.* [9] propose to represent oriented bounding boxes with a set of polar rays, which gets rid of the usage of rotated angles and has a competitive detection effect with five parameters representation. Chen *et al.* [8] propose the PIoU loss to improve the accuracy of rotation angles and rotation IoU, which calculates IoU score in pixel wise and the IoU is continuously differentiable. Despite the improvement on detection accuracy, the aforementioned methods largely rely on the customized layers like proposal predefinition or postprocessing, which are always operated on hundreds of samples. In contrast, the proposed FrequencyFormer builds a simplified paradigm for oriented object detection, which can predict objects at once and is trained end-to-end.

2.2. Transformer in Object Detection

Inspired by the great success of Transformer in natural language processing, numerous efforts have been made to adapt Transformer on various computer vision tasks, such as semantic segmentation[17], visual tracking[18] and object detection[10, 19–21]. In object detection, Carion *et al.* [10] propose the first transformer-based object detector, DETR, which adopts the encoder-decoder framework to form an end-to-end paradigm, and achieves competitive performance with Fast RCNN [3]. Based on DETR, a series of follow-ups are proposed to further optimize the model design. For example, Zhu *et al.* [19] propose Deformable DETR, which utilizes multi-scale deformable attention to replace the cross attention and capitalize multi-scale image features. Wang *et al.* [20] propose AnchorDETR, which draws on the idea of anchor-free method to learn the 2D center point of objects. Liu *et al.* [22] propose DAB DETR, which adopts a 4D reference anchor with more sufficient information. To accelerate the training convergence speed of transformer detector, Li *et al.* [23] propose a denoising training strategy. Based on that, Zhang *et al.* [24] propose a DINO-DETR detector by further introducing a contrast denoising strategy, which shows good performance on small object detection. Our FrequencyFormer is also built upon the Transformer based encoder-decoder like DETR. To adapt this Transformer-based framework to oriented object detection, we introduce a frequency boosted query update strategy to incorporate object queries with frequency vectors, and a dynamic matching strategy to accelerate the model convergence.

2.3. Discrete Cosine Transform

Discrete Cosine Transform (DCT) is widely used in computer vision tasks, which can compress the image to limited memory. In recent years, many deep learning based works have been made to explore the combination of DCT with CNNs. For example, Ehrlich *et al.* [25] introduce DCT into the deep residual network and achieve highly performance on classification task. Xu *et al.* [26] prove the effectiveness of frequency learning for classification, detection and segmentation. Shen *et al.* [27] combine the DCT technology with Mask-RCNN[4] and achieve great performance in instance segmentation task. Similarly, Xue *et al.*[28] propose a frequency domain decomposition network to enhance feature extraction for prohibited item detection in X-ray images. Furthermore, Davari *et al.* [29] utilize scattering wavelets to extract robust features for facial expression recognition. In the medical domain, Wang [30] employs wavelet energy features combined with a neural network to achieve high precision in breast cancer detection. In FrequencyFormer network, we use DCT to transfer oriented objects into frequency vectors, so that the 2D bounding boxes can be compressed into 1D shape encoding. The shape encoding is capable of capturing the discriminativeness among various rotated objects, which are incorporated with Transformer's object queries to achieve robust object detection.

3. The FrequencyFormer Model

3.1. Overall Framework

Given an image $I \in \mathbb{R}^{W\times H\times 3}$, the proposed FrequencyFormer aims to generate the oriented bounding boxes $\{B_i(x_i, y_i, w_i, h_i, \theta_i)\}_{i=1}^{K}$ and the corresponding category labels $\{C_i\}_{i=1}^{K}$ of K objects. The $(x_i, y_i, w_i, h_i, \theta_i)$ indicate the center point coordinates, width, height, and rotation angle of $i-th$ objects, respectively. As illustrated in Fig. 2, the FrequencyFormer is built on a CNN-Transformer architecture. Here, we exploit ResNet101 [31] as backbone and take the output from $5-th$ residual block as image feature, whose resolution

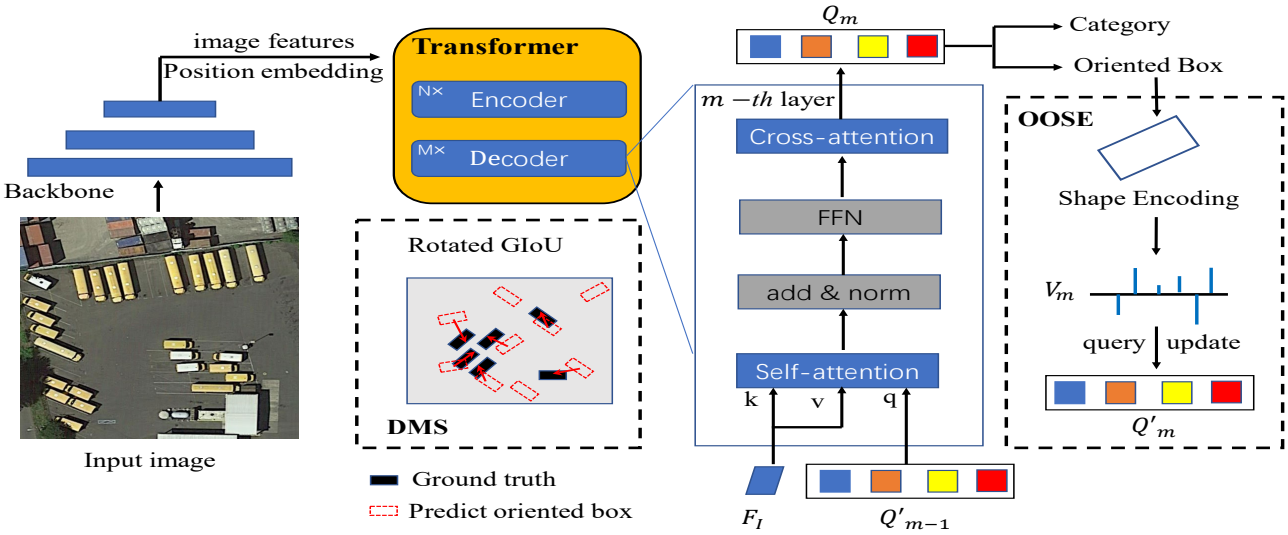

Figure 2. The overall framework of FrequencyFormer. An image $I \in \mathbb{R}^{W\times H\times 3}$ is input to the backbone network to extract image features, with are further combined with position embedding to produce image content F_I by Transformer encoder. In each decoder layer, we encode the output oriented bounding boxes into frequency vectors V_m, and combine them with object queries Q_m to construct updated query features Q'_m, which will be fed to the next decoder layer. In the training phase, a dynamic matching strategy (DMS) is specially designed for the training of FrequencyFormer, which fully considers the matching characteristics of the oriented objects and accelerates the convergence process of the network.

is 1/32 of the input image. It is then fed to the encoder-decoder Transformer for object classification and box regression. The Transformer shares a similar architecture with AnchorDETR [20]. Specifically, the encoder takes the image features and position embedding as input to enhance the interactions of image content by stacking N self-attention layers. The decoder aims to model the interactions between image content (denoted as F_I) and the learned object queries ($Q \in \mathbb{R}^{K\times 256}$) via several cross-attention layers, whose outputs will be fed to a shared feed forward network (FFN) to predict the final detections. We propose a frequency boosted query update strategy to encode effective shape embeddings for object queries to achieve robust oriented object detection. At each decoding layer, the predictions of FFN are first mapped into frequency vectors $V = \{v^i \in \mathbb{R}^{1\times 64}\}_{i=1}^{K}$ by discrete cosine transform, which will be incorporated with the learned object queries to facilitate the object reasoning at next decoding layer. In the training phase, a dynamic matching strategy is specially designed, which fully considers the matching characteristics of oriented objects and accelerates the convergence process of FrequencyFormer.

3.2. Oriented Object Shape Encoding

Previous Transformer based detectors like Conditional DETR[32] and DAB DETR [22] have demonstrated that the object queries are effective at capturing the information of object content and location. To enhance the query features with oriented object shapes, we propose an oriented object shape encoding (OOSE) module. The OOSE is implemented at each decoding layer to capture the shape encoding vectors of intermediate predictions. Specifically, at the $m-th$ decoding layer, we first map the predictions from $(m-1)-th$ layer into frequency domain via discrete cosine transform (DCT). The obtained frequency vectors V will be incorporated with the query features and go through the decoding layer to produce object detections.

Shape encoding. We utilize the DCT technology to convert the 2-D oriented bounding boxes $B \in \mathbb{R}^{K\times 5}$ into the 1-D frequency vectors. As illustrated in Fig. 3, we first crop the enclosed horizontal boxes of the oriented objects to capture their shape priors. Then, the cropped horizontal boxes are converted into binary masks, where the interior pixels of oriented bounding boxes are set to 1 and the remaining part is set to 0. To save the computational costs at multiple decoding layers, we resize the binary mask to a fixed size as 16×16. Finally, the resized mask $mask_r$ are encoded to frequency vectors $V \in \mathbb{R}^{K\times 64}$ via DCT. Through the above process, the shape prior (size, aspect ratio and rotation angle) are implicitly encoded into the finite frequency vectors.

Frequency boosted query update. With the above process of shape encoding, the shape prior of oriented objects are effectively encoded into the frequency vector V. As mentioned in Sec. 1, the oriented objects in different shapes show great variations in frequency

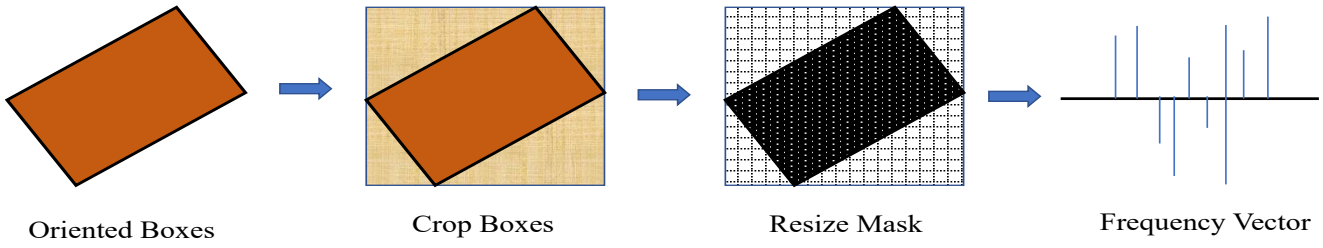

Figure 3. Oriented object shape encoding process. The shape encoding process is divided into three steps: (1) Crop horizontal boxes. (2) Covert to fixed size masks. (3) DCT encoding.

domain. We leverage the obtained frequency vectors to enhance the object queries to capture discriminative shape priors for robust object detection.

In each decoder layer, a shared feed forward network is used to predict oriented bounding boxes by reasoning the cross attention between object queries Q and image content F_I from Transformer encoder. As shown in Fig. 2, the frequency vectors $V_{m-1} \in \mathbb{R}^{K\times 64}$ and the query features $Q_{m-1} \in \mathbb{R}^{K\times 256}$ at the $m-1$ level are concatenated along the channel dimension. Then, the updated query features Q'_{m-1} are obtained by feeding the concatenated features into an MLP layer. The process for query update can be summarized as follows:

$$Q'_{m-1} = MLP(Cat(V_{m-1}, Q_{m-1})) \tag{1}$$

where $Cat()$ indicates the concatenation along channel dimension. The updated queries $Q'_{m-1} \in \mathbb{R}^{K\times 256}$ together with the image content F_I are fed to the $m-th$ Transformer decoder layer and FFN to produce oriented boxes and the corresponding labels. Note that in the first decoder layer, the object queries Q_0 and frequency vectors are initialized with random numbers.

3.3. Dynamic Matching Strategy

The DETR based detectors [10, 12] introduce a bipartite matching mechanism between predictions and ground truth, so that the trained model can generalize well without applying any customized post-processing. In oriented object detection, the arbitrary shapes and distribution of oriented objects and the extra prediction on rotation angle bring difficulties to this matching process. To facilitate the training of FrequencyFormer, we propose a dynamic matching strategy (DMS). Similar with previous methods, the proposed DMS also uses the bipartite matching strategy to match the ground truth and predict oriented objects, while position (center point) matching replaces box sizes matching and the IoU matching weights are adjust adaptively along the training process.

Dynamic matching weights. Specifically, the dynamic matching strategy is constructed based on the matching matrix $M_{K\times G}$, which measures the matching degree between K predicted oriented objects and G ground truth objects. It consists of three parts: categories matching matrix M_C, position matching matrix M_P, and the rotated GIoU matching matrix M_{GIoU_R}, which can be represented as follows:

$$M = W_C * M_C + W_{GIoU_R} * M_{GIoU_R} + W_P * M_P \tag{2}$$

$$M_P = \left\|(x_i, y_i) - (\hat{x_i}, \hat{y_i})\right\|_1 \tag{3}$$

$$W_{GIoU_R} = I_{GIoU_R} * (exp(E_j/E_{total}) - 1) \tag{4}$$

where W_C, W_{GIoU_R} and W_P are the weights of categories matching matrix, rotated GIoU matching matrix, and position matching matrix, respectively. I_{GIoU_R} indicates initial GIoU matching weight. E_j indicates the $j-th$ training epoch and the E_{total} is the number of total epochs. M_P only considers the center position of the oriented object bounding boxes and M_C is same to AnchorDETR [20].

The W_C and W_P keep constant in the training process but the rotated GIoU matching weights W_{GIoU_R} is adjusted adaptively as the training process goes on. As illustrated in Equation 4, the W_{GIoU_R} gradually increases from 0 with the training progress. With such dynamic strategy, the FrequencyFormer pays more attention to the position regression and classification at the early stage of training, and focuses more on oriented object boxes shapes at the later stage.

Rotated GIoU matching. IoU matching is an critical component in the matching strategy. Different from the general detector that calculate IoU on horizontal bounding boxes, the FrequencyFormer adopts the rotated Generalized Intersection over Union (GIoU) [33] to measure the IoU matching similarity. The rotated GIoU considers not only the IoU between two oriented objects, but their distance over rotated boxes, which could match the appropriate ground truth and help the network converge smoothly. In bipartite matching strategy, the smaller values in the matching matrix indicate that the corresponding samples are more matched with the ground truth. So we convert

the rotated GIoU to a rotated GIoU matching matrix M_{GIoU_R} by:

$$M_{GIoU_R} = 1 - GIoU_R \tag{5}$$

$$GIoU_R = IoU_R - \frac{Area/(A \cup A')}{|Aera|} \tag{6}$$

where A and A' are two oriented bounding boxes, the $Area$ is the smallest convex shapes enclosing both A and A'.

4. Experiments

4.1. Dataset and Metrics

Datasets. To evaluate the effectiveness of FrequencyFormer, we conduct a series of experiments on two public datasets, *i.e.*, DOTA dataset[14] and HRSC2016 dataset [15].

DOTA dataset is widely used in aerial images for object detection. The dataset is split into three parts: training set (1/2), validation set (1/3) and test set (1/6). The images in DOTA dataset have a large resolution at 4000×4000 and are annotated by two types: horizontal bounding boxes and oriented bounding boxes. The DOTA dataset includes 15 categories, including plane (PL), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (HA), bridge (beidge), large vehicle (LV), small vehicle (SV), helicopter (HC), roundabout (RA), soccer ball field (SBF) and swimming pool (SP).

HRSC2016 dataset is specially captured for ship detection with oriented bounding box annotation. The HRSC2016 dataset has 1070 images in total with a resolution at 1000×600. The images are split into training set (436 images), validation set (181 images) and test set (444 images).

Metrics. To make a fair comparison with other methods, we use mean Average Precision (mAP) to evaluate the detection accuracy. The mAP is related to the Precision and Recall, which are defined as

$$Precision = TP/(TP + FP) \tag{7}$$

$$Recall = TP/(TP + FN) \tag{8}$$

where TP, FP and FN are the number of true positive samples, false positive samples and false negative samples, respectively.

4.2. Implementation Details

Training. The proposed FrequencyFormer is trained on 4 NVIDIA Tesla P40 GPUs with batch size as 4 and the initial learning rate as 1e-4. We train the network for 50 epochs and the learning rate decays 10 times at 40 epochs. Multi-scales ($0.5\times, 1.0\times, 1.5\times$) training is adopted as data augmentation during training. We adopt multiple loss functions to supervise the FrequencyFormer. We utilize the rotated GIoU loss together with L1 loss to supervise the oriented bounding boxes regression and focal loss for object classification. The corresponding weights are set to 2, 5 and 2, respectively. In the matching stage, the initial weights I_{GIoU_R} of rotated GIoU matching matrix is set to 2, the weights of categories matching matrix W_C and position matching matrix W_P are set to 2 and 5, respectively. For the FrequencyFormer network structure, the encoder layers number M and decoder layers number N are all set to 6. The number of object queries K is set to 300.

Inference. In the inference phase, the score threshold is set to 0.1 to distill the final detection from 300 samples. When testing on DOTA dataset, we first crop the images into local patches at 1024×1024 with an overlap of 200 pixels. Then the final results for each image are obtained by merging the predictions of cropped images by rotated NMS with threshold as 0.3.

4.3. Comparison with State-of-Arts Methods

DOTA dataset. We compare the FrequencyFormer with other state-of-the-art methods for oriented object detection on DOTA test set. As illustrated in Tab.1, we list the results of previous oriented object detectors and our method on DOTA test set. As we can see, the anchor-based methods have a high detection accuracy with 80.87% on mAP and the anchor-free methods achieves the 77.63% on mAP. Our FrequencyFormer achieves 76.00% on mAP, which has a comparable performance with convolutional oriented object detectors. Besides, to intuitively show the of detection accuracy of FrequencyFormer, we visualize the oriented object detection results in Fig. 5. As illustrated in Fig. 5, our FrequencyFormer can accurately detect the objects with arbitrary shapes and distribution.

HRSC2016 dataset. We also compare our FrequencyFormer network with other methods on HRSC2016 test set. As illustrated in Tab. 2, the existing oriented object detectors are all convolutional-style detectors, they gets the 90.5% mAP in VOC2007 metrics and 97.26% mAP in VOC2012 metrics, our FrequencyFormer network achieves 90.36% mAP in VOC2007 metrics and 96.91% mAP in VOC2012 metrics. The quantitative comparison results demonstrate that our FrequencyFormer network has a satisfactory detection effect on oriented object detection in HRSC2016 dataset. Besides, we give some visual results in Fig. 7, which demonstrate that the FrequencyFormer can accurately detect the ships with arbitrary orientations.

4.4. Ablation Studies

To evaluate the effectiveness of each module in FrequencyFormer, we conduct a series of ablation

Table 1. Quantitative comparisons between our FrequencyFormer and other methods on DOTA test set.

Method	mAP(%)	PL	BD	Bridge	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
Anchor-based method																
RoI-Transformer[6]	69.56	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67
CAD-Net[34]	69.9	87.8	82.4	49.4	73.5	71.1	63.5	76.7	**90.9**	79.2	73.3	48.4	60.9	62.0	67.0	62.2
SCRDet[35]	72.61	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	86.86	65.02	66.68	66.25	68.24	65.21
R^3Det[36]	72.81	89.24	80.81	51.11	65.62	70.67	76.03	78.32	90.83	84.89	84.42	65.10	57.18	68.10	68.98	60.88
Gliding Vertex[37]	75.02	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	**70.91**	72.94	70.86	57.32
APE[38]	75.75	89.96	83.62	53.42	76.03	74.01	77.16	79.45	90.83	87.15	84.51	67.72	60.33	74.61	71.84	65.55
CSL[39]	76.17	**90.25**	**85.53**	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93
DCL[5]	77.37	89.26	83.60	53.54	72.76	79.04	82.56	87.31	90.67	86.59	86.98	67.49	66.88	73.29	70.56	69.99
ReDet[40]	80.10	88.81	82.48	60.83	80.82	78.34	86.06	88.31	90.87	88.77	87.03	68.65	66.90	79.26	79.71	74.67
Oriented RCNN[41]	**80.87**	89.84	85.43	**61.09**	**79.82**	**79.71**	**85.35**	**88.82**	90.88	86.68	**87.73**	**72.21**	70.80	**82.42**	**78.18**	**74.11**
Anchor-free method																
Axis Learning[42]	65.98	79.53	77.15	38.59	61.15	67.53	70.49	76.30	89.66	79.07	83.53	47.27	61.01	56.28	66.06	36.05
O2-DNet[43]	71.04	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03
DRN[44]	73.23	**89.71**	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48
CBDA-Net[7]	75.74	89.17	**85.92**	50.28	65.02	77.72	82.32	**87.89**	90.48	86.47	85.90	**66.85**	66.48	67.41	71.33	62.89
PRNet[9]	76.50	88.78	85.51	51.89	73.42	74.36	77.98	87.78	**90.86**	86.50	**86.84**	60.69	**69.84**	75.29	72.22	**67.22**
Oriented RepPoints[45]	**77.63**	89.11	82.32	**56.71**	**74.95**	**80.70**	**83.73**	87.67	90.81	**87.11**	85.85	63.60	68.60	**75.95**	**73.54**	63.76
Transformer-style method																
O^2DETR[46]	72.15	86.01	75.92	46.02	66.65	**79.70**	**79.93**	**89.17**	90.44	81.19	76.00	56.91	62.45	64.22	65.80	58.96
Ours	**76.00**	**88.90**	**81.79**	**52.02**	**76.66**	74.50	77.29	88.00	**90.76**	**86.08**	**87.25**	**62.22**	**66.92**	**74.38**	**72.57**	**60.65**

Table 2. Quantitative comparisons between our method and other methods for oriented object detection task on HRSC2016 test set. $*$ indicates that the result is evaluated under VOC2012 metrics, while other methods are all evaluated under VOC2007 metrics. w. T indicates whether the corresponding method use transformer structure.

Method	w. T	Backbone	mAP(%)
RoI-Transformer[6]	×	ResNet101	86.2
R^3Det[36]	×	MobileNetV2	86.7
PIoU[8]	×	DLA-34	89.2
DRN[44]	×	Hourglass-104	92.7*
Oriented RCNN[43]	×	ResNet50-FPN	**90.5**/ 96.7*
Oriented RepPoints[45]	×	ResNet50-FPN	90.38/**97.26***
Ours	✓	ResNet101-DC5	90.36/96.91*

Table 3. Ablation experiments results on DOTA validation set.

Methods	OOSE	DMS	mAP(%)
Baseline			57.71
Our method	✓		58.98(**+1.27**)
		✓	59.56(**+1.85**)
	✓	✓	60.80(**+3.09**)

experiments on DOTA validation set. We modify the AnchorDETR detector [20] to oriented object detection as a baseline model, by extending the output coordinates with 5 dimensions (*i.e.* x, y, width, height, rotation angle).

Effectiveness of each module. As illustrated in Tab. 3, the baseline model achieves 57.71% mAP on DOTA validation set. After the baseline model uses the OOSE module to update the query features, the baseline+OOSE model achieves 58.98% mAP. After the baseline model adopt the DMS module to accelerate the training process, the baseline+DMS model achieves 59.56% mAP. After we add the OOSE module together with DMS module on the baseline model, the baseline+OOSE+DMS model achieves 60.80% mAP. These results demonstrate that both the OOSE module and DMS module have improvement effect on the oriented object detection task.

Analysis of DMS. To verify the effectiveness of dynamic training strategy, we compare the AP results with the training process between baseline model and baseline+DMS model on DOTA validation set. As illustrated in Fig. 4, the AP of baseline+DMS model is higher than the baseline model during the whole training process. Specially, the baseline+DMS model has faster coverage speed than the baseline model in initial stage. The results demonstrate that the DMS is effective for accelerating the convergence speed of network and improving oriented object detection accuracy.

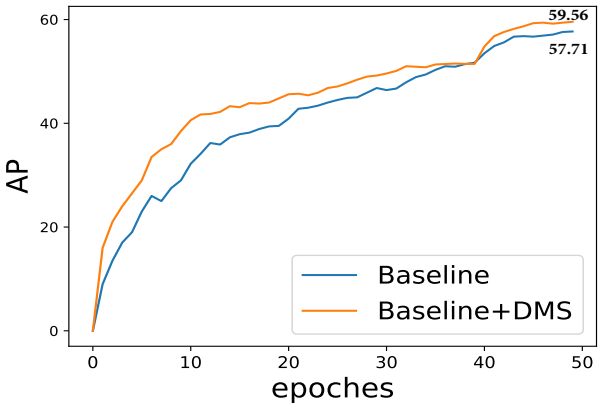

Figure 4. Comparison of AP with the training process between baseline and baseline+DMS models on DOTA validation set.

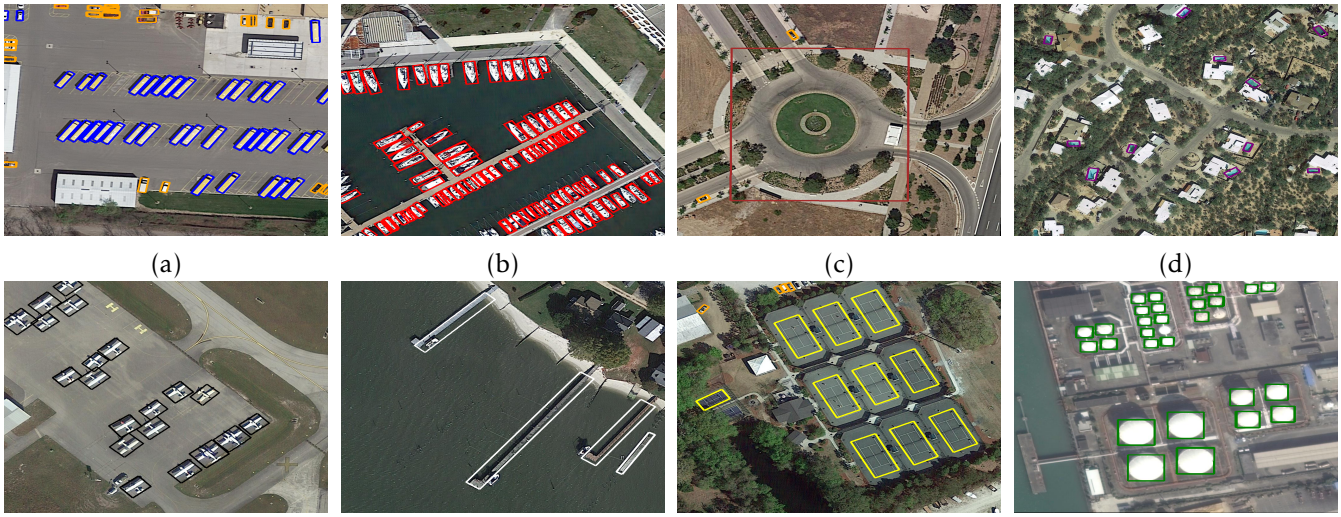

Figure 5. Examples of detection results on the DOTA test set. (a) Large vehicle and small vehicle. (b) Ship. (c) Roundabout. (d) Swimming pool. (e) Plane. (f) Harbor. (g) Tennis court. (h) Storage tank.

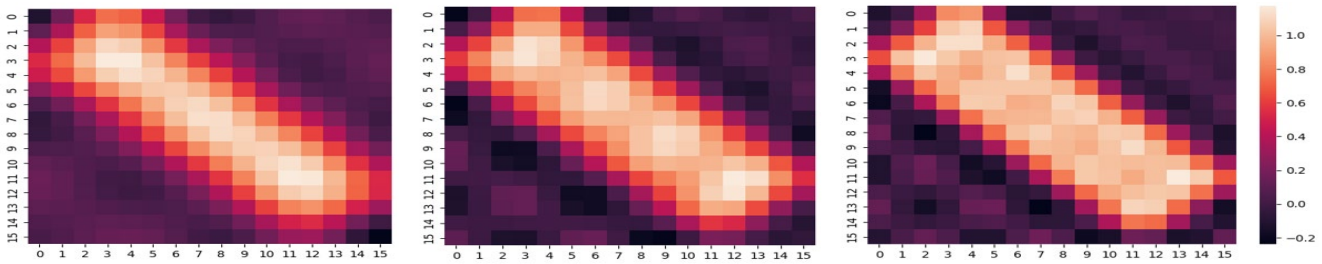

Figure 6. Visualization decoding results of frequency with different dimensions. We present the decoding visualization results in different frequency dimensions (32, 64, 96).

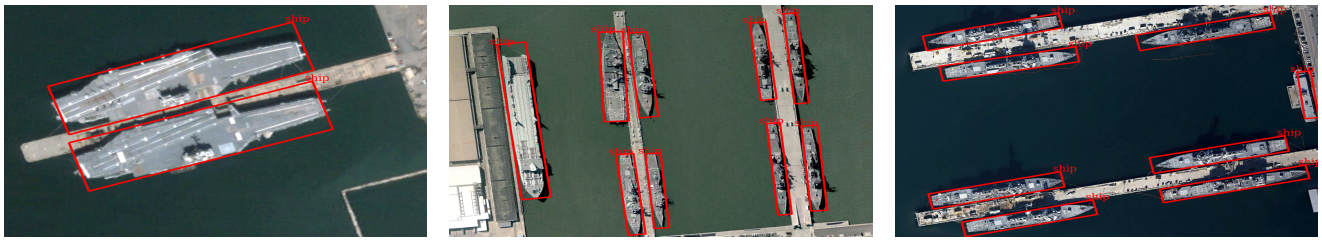

Figure 7. Visual examples of detection results on the HRSC2016 test set.

Analysis of frequency dimensions. In OOSE module, we encode the oriented bounding boxes into frequency vectors by DCT. To intuitively present the contained information in different frequency dimensions, we decode frequency vectors with different dimensions and visualize the results. As illustrated in Fig. 6, from left to right, there are images from the frequency decoding with dimensions of 32, 64 and 96, respectively. With the increase of dimensions, the shape contour of oriented objects becomes more obvious and the object shape is clearer. The frequency vectors decoding visualization results demonstrate that the larger the dimension of frequency vector, the more information it contains.

Besides, to explore the influence of shape encoding frequency dimension on detection accuracy, we conduct different dimensions of frequency and compare the results on mAP. As illustrated in Tab. 4, we report the mAP of different dimensions (32, 64, 96) on DOTA validation set, the mAP gets 58.87%, 58.98% and 58.99% on frequency dimensions 32, 64 and 96, respectively. The results demonstrate that the detection accuracy increases with the increase of shape encoding frequency dimensions.

Table 4. Studies of different frequency dimensions. We report the mAP of different dimensions (32, 64, 96) on DOTA validation set.

Dimensions	32	64	96
mAP(%)	58.87	58.98	58.99

5. Limitations

The proposed FrequencyFormer belongs to transformer-style detector, which has higher detection accuracy but weakness in longer inference time compared with convolution neural networks. In the transformer decoder layer, the OOSE module uses the DCT technology to convert the shape information of oriented bounding boxes into frequency information, which will increase the computational load and complexity of the detection model undoubtedly. The FrequencyFormer model is an effective attempt to introduce frequency into the transformer structure, which provides a new design idea for the oriented object detector. In the future, we will explore more appropriate structures and introduce frequency information into the transformer structure, and reduce the computational complexity.

6. Conclusion

This paper proposes a novel transformer-style oriented object detector, FrequencyFormer. The FrequencyFormer encodes the oriented boxes into frequency vectors to enhance the shape priors for object queries. Besides, we propose a dynamic matching strategy to accelerate the network convergence, where the matching weights are adjusted adaptively as the training goes on. Experimental results on DOTA dataset and HRSC2016 dataset demonstrate that the proposed FrequencyFormer can accurately detect oriented objects and achieves competitive performance with other state-of-the-arts methods, without applying any customized post-processing.

Acknowledgements. The paper is supported by Hebei Province Natural Science Foundation Funded Project (F2024202033), National Funded Postdoctoral Research Program (GZC20241900), Natural Science Foundation Program of Xinjiang Uygur Autonomous Region (2024D01A141), Tianchi Talents Program of Xinjiang Uygur Autonomous Region, and the open project of Dazhou Key Laboratory of Government Data Security (ZSAQ202502).

References

[1] Pu, X. and Xu, F. (2024) Low-rank adaption on transformer-based oriented object detector for satellite onboard processing of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* .

[2] Su, Y., Chen, Z., Du, Y., Ji, Z., Hu, K., Bai, J. and Gao, X. (2024) Explicit relational reasoning network for scene text detection : 7069–7077.

[3] Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28**.

[4] He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017) Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*: 2961–2969.

[5] Yang, X., Hou, L., Zhou, Y., Wang, W. and Yan, J. (2021) Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 15819–15829.

[6] Ding, J., Xue, N., Long, Y., Xia, G.S. and Lu, Q. (2019) Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 2849–2858.

[7] Liu, S., Zhang, L., Lu, H. and He, Y. (2021) Center-boundary dual attention for oriented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **60**: 1–14.

[8] Chen, Z., Chen, K., Lin, W., See, J., Yu, H., Ke, Y. and Yang, C. (2020) Piou loss: Towards accurate oriented object detection in complex environments. In *European Conference on Computer Vision* (Springer): 195–211.

[9] Liu, S., Zhang, L., Hao, S., Lu, H. and He, Y. (2021) Polar ray: A single-stage angle-free detector for oriented object detection in aerial images. In *Proceedings of the 29th ACM International Conference on Multimedia*: 3124–3132.

[10] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020) End-to-end object detection with transformers. In *European conference on computer vision* (Springer): 213–229.

[11] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C. and Wang, Y. (2021) Transformer in transformer. *Advances in Neural Information Processing Systems* **34**.

[12] Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J. (2020) Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* .

[13] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X. and Lu, H. (2021) Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 8126–8135.

[14] Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M. *et al.* (2018) Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 3974–3983.

[15] Liu, Z., Yuan, L., Weng, L. and Yang, Y. (2017) A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods* (SciTePress), **2**: 324–331.

[16] Sheng, T., Chen, J. and Lian, Z. (2021) Centripetaltext: An efficient text instance representation for scene text detection. *Advances in Neural Information Processing Systems* **34**.

[17] Strudel, R., Garcia, R., Laptev, I. and Schmid, C. (2021) Segmenter: Transformer for semantic segmentation. In

Proceedings of the IEEE/CVF International Conference on Computer Vision: 7262–7272.

[18] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X. and Lu, H. (2021) Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 8126–8135.

[19] Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J. (2020) Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*.

[20] Wang, Y., Zhang, X., Yang, T. and Sun, J. (2021) Anchor detr: Query design for transformer-based detector. *arXiv preprint arXiv:2109.07107* .

[21] Ngo, D.D., Vo, V.L., Le, M.H., Hoc-Phan and Nguyen, M.H. (2025) Transformer based ship detector: An improvement on feature map and tiny training set. *EAI Endorsed Transactions on Industrial Networks Intelligent Systems* **12**(1).

[22] Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J. *et al.* (2021) Dab-detr: Dynamic anchor boxes are better queries for detr. In *International Conference on Learning Representations*.

[23] Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M. and Zhang, L. (2022) Dn-detr: Accelerate detr training by introducing query denoising. *arXiv preprint arXiv:2203.01305* .

[24] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M. *et al.* (2022) Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605* .

[25] Ehrlich, M. and Davis, L.S. (2019) Deep residual learning in the jpeg transform domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*: 3484–3493.

[26] Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.K. and Ren, F. (2020) Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 1740–1749.

[27] Shen, X., Yang, J., Wei, C., Deng, B., Huang, J., Hua, X.S., Cheng, X. *et al.* (2021) Dct-mask: Discrete cosine transform mask representation for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 8720–8729.

[28] Xue, Z., Wang, B., Xie, Y., Li, Z., Fan, X., Lin, C., Wei, P. *et al.* (2025) Fdd-yolo: A lightweight multi-scale prohibited items detection model. *EAI Endorsed Transactions on AI and Robotics* **4**(1).

[29] Davari, M., Harooni, A., Nasr, A. and Savoji, K. (2024) Improving recognition accuracy for facial expressions using scattering wavelet. *EAI Endorsed Transactions on AI and Robotics* **3**(1).

[30] Wang, J. (2023) Breast cancer detection via wavelet energy and feed-forward neural network trained by genetic algorithm. *EAI Endorsed Transactions on AI and Robotics* **2**(1).

[31] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 770–778.

[32] Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L. *et al.* (2021) Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*: 3651–3660.

[33] Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y. and Yang, R. (2019) Iou loss for 2d/3d object detection. In *2019 International Conference on 3D Vision (3DV)* (IEEE): 85–94.

[34] Zhang, G., Lu, S. and Zhang, W. (2019) Cad-net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **57**(12): 10015–10024.

[35] Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X. *et al.* (2019) Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*: 8232–8241.

[36] Yang, X., Liu, Q., Yan, J., Li, A., Zhang, Z. and Yu, G. (2019) R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612* **2**(4): 2.

[37] Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.S. and Bai, X. (2020) Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE transactions on pattern analysis and machine intelligence* **43**(4): 1452–1459.

[38] Zhu, Y., Du, J. and Wu, X. (2020) Adaptive period embedding for representing oriented objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing* **58**(10): 7247–7257.

[39] Yang, X. and Yan, J. (2020) Arbitrary-oriented object detection with circular smooth label. In *European Conference on Computer Vision* (Springer): 677–694.

[40] Han, J., Ding, J., Xue, N. and Xia, G.S. (2021) Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 2786–2795.

[41] Xie, X., Cheng, G., Wang, J., Yao, X. and Han, J. (2021) Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*: 3520–3529.

[42] Xiao, Z., Qian, L., Shao, W., Tan, X. and Wang, K. (2020) Axis learning for orientated objects detection in aerial images. *Remote Sensing* **12**(6): 908.

[43] Wei, H., Zhang, Y., Chang, Z., Li, H., Wang, H. and Sun, X. (2020) Oriented objects as pairs of middle lines. *ISPRS Journal of Photogrammetry and Remote Sensing* **169**: 268–279.

[44] Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C. *et al.* (2020) Dynamic refinement network for oriented and densely packed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 11207–11216.

[45] Li, W. and Zhu, J. (2021) Oriented reppoints for aerial object detection. *arXiv preprint arXiv:2105.11111* .

[46] Ma, T., Mao, M., Zheng, H., Gao, P., Wang, X., Han, S., Ding, E. *et al.* (2021) Oriented object detection with transformer. *arXiv preprint arXiv:2106.03146* .