Enhanced Oriented Object Detection in Remote Sensing Using Deformable Quadrilateral Attention and PCA

Li-Dan Kuang, Junwu Xie, Yan Gui, Jianming Zhang, Xiaoyong Tang
School of Computer Science
Changsha University of Science and Technology
Changsha, China

***Abstract*—Current oriented object detection methods often underutilize the distinctive properties of rotated bounding boxes. To bridge this gap, we introduce a novel approach centered on a learnable Deformable Quadrilateral Attention (DQA) mechanism, explicitly tailored for oriented detection. DQA generalizes window-based attention by dynamically adapting to a flexible quadrilateral formulation. An end-to-end regression module predicts a transformation matrix that deforms a default window into an oriented quadrilateral, enabling adaptive token sampling and enriched context modeling for variably shaped and rotated objects. Furthermore, we integrate Principal Component Analysis (PCA) into the detection head to efficiently deduce rotated bounding boxes from the predicted quadrilateral, enhancing both accuracy and efficiency. Combined, our DQA head with PCA achieves state-of-the-art results: 98.90% mAP on HRSC2016, 79.38% mAP on DOTA-v1.0, and 67.24% mAP on DIOR-R, demonstrating its strong capability for oriented object detection.**

***Index Terms*—Oriented Object Detection, Remote Sensing, Deformable Attention, Vision Transformer, Principal Component Analysis.**

I. Introduction

Remote sensing imagery presents unique challenges for object detection, where targets appear in arbitrary orientations and exhibit diverse aspect ratios. While recent advances in oriented object detection have improved performance through enhanced backbones and specialized detection heads, most methods still rely on static, axis-aligned window designs or treat orientation as a secondary regression parameter. This limits their ability to fully capture the geometric characteristics of rotated objects.

A promising direction is explored by Spatial Transform Decoupling (STD), which decouples bounding box attributes using separate Transformer branches. However, STD's dependence on rigid rectangular windows constrains its flexibility in modeling irregular object layouts commonly found in aerial scenes. To overcome this, we propose a novel detection head incorporating a Deformable Quadrilateral Attention (DQA) mechanism. DQA generalizes window attention by allowing the receptive field to adaptively deform into a data-driven quadrilateral, better aligning with the actual shape and orientation of the target.

Corresponding author: Li-Dan Kuang (kuangld@csust.edu.cn).

Our core contribution is threefold:

1) We propose the Deformable Quadrilateral Attention (DQA), a learnable attention mechanism that dynamically adjusts the sampling window into an oriented quadrilateral, facilitating more precise feature extraction for rotated objects.
2) We incorporate Principal Component Analysis (PCA) into the detection head to efficiently derive the principal orientation of the predicted quadrilateral, simplifying the process of obtaining accurate rotated bounding boxes.
3) To the best of our knowledge, this work is the first to introduce a data-driven adaptive window mechanism into remote sensing object detection, representing a significant deviation from traditional static window designs.

Extensive experiments validate our approach, achieving new state-of-the-art results on HRSC2016 (98.90% mAP), DOTA-v1.0 (79.38% mAP), and DIOR-R (67.24% mAP). These results underscore the effectiveness of tailoring attention mechanisms to the inherent geometry of oriented objects.

II. Related Work

A. Remote Sensing Object Detection Frameworks

Rotated object detection extends traditional horizontal detection frameworks by incorporating the representation of oriented bounding boxes. In aerial imagery, objects typically have arbitrary orientations and dense distributions, with significant aspect ratio variations. To address these challenges, researchers have proposed several two-stage detection methods, such as ICN [9]and ROI Transformer [3], which inherit the Faster R-CNN framework. These methods generate oriented proposals through Region Proposal Networks and perform fine regression in the detection head. On the other hand, one-stage methods, such as DRN [10] and RSDet [11], have gained attention due to their faster detection speeds.

B. Vision Transformers and Attention Mechanisms

Although ViTs [14] have demonstrated excellent performance in various vision tasks, ViTs still face challenges when

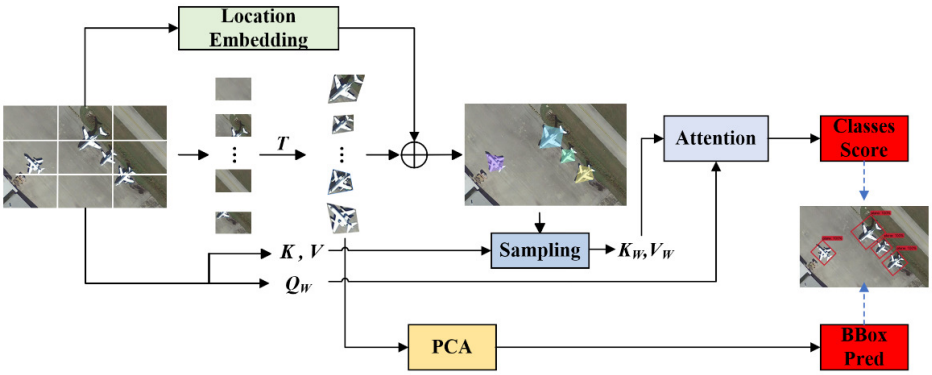

Fig. 1: The framework of the proposed Deformable Quadrilateral Attention (DQA) method.

handling oriented objects. Recently, RVSA [15] has made preliminary attempts to solve the rotated object detection problem by improving the window attention mechanism. STD introduced independent Transformer branches in the detection head to separately predict the position, size, and angle of bounding boxes. However, existing attention mechanisms have not been optimized for the rotational characteristics of irregular quadrilaterals, leading to limitations when detecting targets such as airplanes and ships.

Currently, most existing rotated object detection methods add angle information to a horizontal bounding box. However, this approach can lead to errors in detecting irregular quadrilateral objects, such as airplanes, where the horizontal box is incorrect, consequently resulting in inaccurate rotated bounding box predictions. To address this issue, we propose a novel detection head that first generates predictions for irregular quadrilaterals. Then, PCA is applied to determine the primary orientation, which is used to refine the rotated bounding box information. Meanwhile, a separate branch is employed to predict the object's class, utilizing a divide-and-conquer strategy to accurately perform remote sensing object detection tasks.

To address this issue, this paper proposes a deformable quadrilateral attention mechanism, which uses a data-driven window design to better adapt to the geometric characteristics of rotated objects.

III. Method

A. Deformable Quadrilateral Attention (DQA) Mechanism

Existing attention mechanisms predominantly rely on fixed, axis-aligned windows, which lack the geometric flexibility to effectively represent oriented or irregularly shaped objects. To transcend this limitation, we introduce a Deformable Quadrilateral Attention (DQA) mechanism—a fully learnable, data-driven approach that dynamically adapts the spatial support of each attention window. As shown in Fig. 1, DQA parameterizes each window as a projective quadrilateral, enabling adaptive adjustment of position, scale, orientation, aspect ratio, and perspective, thereby providing a more precise geometric prior for object representation.

1) Window Generation: Given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels, respectively, we first partition it into non-overlapping base windows of size $w \times w$. For each window, we project the features into query (Q), key (K), and value (V) tokens via linear transformations:

$$Q, K, V = \text{Linear}(X), \tag{1}$$

where $\text{Linear}(\cdot)$ represents a learnable linear transformation applied to X. The query tokens are directly used for the DQA computation, while the K and V tokens are reshaped into 2D feature maps to facilitate subsequent differentiable sampling.

2) Quadrilateral Generation via Learned Projective Transformation: The core of DQA is a lightweight quadrilateral prediction module that estimates a set of transformation parameters for each base window. These parameters define a projective transformation T, which maps the regular rectangular window to an arbitrary quadrilateral in the feature domain, thereby dynamically modeling the geometric attributes of objects based on visual content.

For a base window feature $X_w \in \mathbb{R}^{w\times w\times C}$, the module first condenses spatial information via average pooling, applies a LeakyReLU non-linearity, and outputs a 9-dimensional proxy vector t via a 1×1 convolution:

$$t = \text{Conv}_{1\times 1}\big(\text{LeakyReLU}(\text{AvgPool}(X_w))\big). \tag{2}$$

The vector t is then decoded into a sequence of elementary transformations: scaling (T_s), shear (T_h), rotation (T_r), translation (T_t), and projection (T_p). Their matrices are constructed

as:

$$T_s = \begin{pmatrix} t_1+1 & 0 & 0 \\ 0 & t_2+1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad T_h = \begin{pmatrix} 1 & t_3 & 0 \\ t_4 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$T_r = \begin{pmatrix} \cos t_5 & -\sin t_5 & 0 \\ \sin t_5 & \cos t_5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad T_t = \begin{pmatrix} 1 & 0 & \beta_1 t_6 \\ 0 & 1 & \beta_2 t_7 \\ 0 & 0 & 1 \end{pmatrix},$$

$$T_p = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t_8 & t_9 & 1 \end{pmatrix}, \tag{3}$$

where $\beta_1 = \frac{W}{w}$ and $\beta_2 = \frac{H}{w}$ scale the translation to accommodate varying input resolutions. The full projective transformation is composed as:

$$T = T_s \cdot T_h \cdot T_r \cdot T_t \cdot T_p. \tag{4}$$

The matrices are composed in a fixed order (scale, shear, rotate, translate, then project) to construct a general projective transformation. For a point (x, y) in the base window (in homogeneous coordinates $\begin{bmatrix} x, y, 1 \end{bmatrix}^\top$), its projected location is computed as:

$$\begin{bmatrix} x_{\mathrm{t}} \\ y_{\mathrm{t}} \\ z_{\mathrm{t}} \end{bmatrix} = T \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad (x', y') = \left(\frac{x_{\mathrm{t}}}{z_{\mathrm{t}}}, \frac{y_{\mathrm{t}}}{z_{\mathrm{t}}} \right), \tag{5}$$

where $\begin{bmatrix} x_t, y_t, z_t \end{bmatrix}^\top$ denotes the intermediate homogeneous coordinates after projective transformation by T, and (x', y') denotes the final 2D coordinates of the point within the target quadrilateral, which are subsequently used for bilinear sampling of the key and value features. Applying this transformation to all points in parallel yields the target quadrilateral region. Notably, standard window attention is a special case of DQA when $t = 0$ (i.e., T is the identity matrix). We initialize the prediction module to produce near-zero parameters, ensuring stable training onset.

3) Differentiable Sampling and Attention Computation: Given the coordinates of the projected quadrilateral, we sample the corresponding key and value tokens from the K and V feature maps using bilinear interpolation, obtaining window-specific tokens K_w and V_w. The preceding query tokens are directly used in the DQA computation(i.e., $Q_w = Q$). To handle projections that extend beyond the feature map boundaries, we employ a boundary-aware sampling strategy: coordinates inside the valid region are interpolated normally, while those outside are assigned a zero vector. This preserves gradient flow without introducing boundary artifacts.

The attention output for the window is then computed as:

$$F = \Big(\mathrm{softmax}\big(Q_w K_w^\top\big) + r \Big) V_w, \tag{6}$$

where r denotes the learnable relative position bias encoding spatial relationships within the quadrilateral.

4) Regularization for Boundary Awareness: To discourage excessive projection outside the feature map—which would dilute feature information—we introduce a piecewise linear regularization loss on the token coordinates $(x_{q,i}, y_{q,i})$, which are normalized to the range [-1,1]:

$$\mathcal{R}(x) = \begin{cases} -\lambda, & x < -1 \\ 0, & -1 \le x \le 1 \\ \lambda, & x > 1 \end{cases} \tag{7}$$

$$L_{reg} = \sum_i \left[\mathcal{R}(x_{q,i}) + \mathcal{R}(y_{q,i}) \right], \tag{8}$$

where λ is a hyperparameter that controls the penalty strength. This loss is added to the main task objective, gently biasing quadrilaterals to remain within the spatial extent of the feature map while retaining full flexibility.

5) Implementation Details: During training, quadrilaterals are allowed to overlap freely to strengthen cross-window interaction, and attention masks are used to normalize weights and avoid duplicate counting. For out-of-bounds areas, the zero-padding strategy ensures gradient stability. The entire pipeline—parameter prediction, coordinate transformation, sampling, and attention—is differentiable, enabling end-to-end optimization.

DQA generalizes standard window attention and provides a unified, learnable framework for geometry-aware feature aggregation, effectively improving the detection accuracy of rotated and irregular objects.

B. Principal Component Analysis (PCA)

In oriented object detection, targets exhibit diverse geometries (orientation, shape, size) that are poorly captured by traditional rectangular attention mechanisms. To address this, we employ Principal Component Analysis (PCA) on predicted quadrilateral vertices to determine the target's principal direction, enhancing detection accuracy.

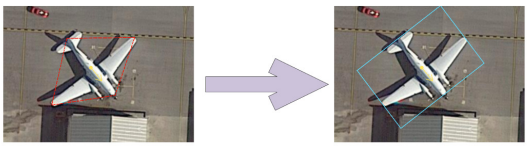

Fig. 2: Illustration of the PCA effect. Left subfigure is the BBox after obtaining the keypoint coordinates, and right subfigure is the BBox after applying PCA.

Given vertices $A(x_1, y_1)$, $B(x_2, y_2)$, $C(x_3, y_3)$, and $D(x_4, y_4)$, we first center the data by subtracting the centroid $(\bar{x}, \bar{y})$:

$$X' = \begin{bmatrix} x_1 - \bar{x}, & y_1 - \bar{y} \\ x_2 - \bar{x}, & y_2 - \bar{y} \\ x_3 - \bar{x}, & y_3 - \bar{y} \\ x_4 - \bar{x}, & y_4 - \bar{y} \end{bmatrix}, \tag{9}$$

TABLE I: Experimental results on DOTA-v1.0 dataset [7] under single-scale training and testing setting. ResNet-50 [16] and LSKNet-S [2] is pretrained on ImageNet-1K for 300 epochs similar to the compared methods [17], [20].

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
DETR-based																	
AO^2-DETR [21]	ResNet-50	87.99	79.46	45.74	66.64	78.90	73.90	73.30	90.40	80.55	85.89	55.19	63.62	51.83	70.15	60.04	70.91
ARS-DETR [22]	ResNet-50	86.61	77.26	48.84	66.76	78.38	78.96	87.40	90.61	82.76	82.19	54.02	62.61	72.64	72.80	64.96	73.79
One-stage																	
Rotated FCOS [23]	ResNet-50	88.52	77.54	47.06	63.78	80.42	80.50	87.34	90.39	77.83	84.13	55.45	65.84	66.02	72.77	49.17	72.45
R3Det	ResNet-50	89.00	75.60	46.64	67.09	76.18	73.40	79.02	90.88	78.62	84.88	59.00	61.16	63.65	62.39	37.94	69.70
S^{2}ANet [4]	ResNet-50	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
	ARC-R50 [13]	89.28	78.77	53.00	72.44	79.81	77.84	86.81	90.88	84.27	86.20	60.74	68.97	66.35	71.25	65.77	75.49
Two-stage																	
ReDet [12]	ResNet-50	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59	76.25
ROI Transformer	ResNet-50	89.01	77.48	51.64	72.07	74.43	77.55	87.76	90.81	79.71	85.27	58.36	64.11	76.50	71.99	54.06	74.05
Rotated Faster	ResNet-50	89.40	81.81	47.28	67.44	73.96	73.12	85.03	90.90	85.15	84.90	56.60	64.77	64.70	70.28	62.22	73.17
R-CNN [19]	ARC-R50	89.49	82.11	51.02	70.38	79.07	75.06	86.18	90.91	84.23	86.41	56.10	69.42	65.87	71.90	63.47	74.77
O-RCNN	ResNet-50	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.87
	ARC-R50	89.40	82.48	55.33	73.88	79.37	84.05	88.06	90.90	86.44	84.83	63.63	**70.32**	74.29	71.91	65.43	77.35
	LSKNet-S	89.66	85.52	57.72	75.70	74.95	78.69	**88.24**	90.88	86.79	86.38	66.92	63.77	77.77	**74.47**	64.82	77.49
	PKINet-S [1]	**89.72**	84.20	55.81	**77.63**	**80.25**	84.45	88.12	90.88	**87.57**	86.07	66.86	70.23	77.47	73.62	62.94	78.39
DQA	LSKNet-S	89.71	**85.09**	**57.29**	72.28	**80.28**	**85.64**	88.12	**90.98**	86.56	**86.49**	**67.88**	69.85	**78.89**	73.59	**77.38**	**79.38**

The sample covariance matrix C is then computed from the centered data X' :

$$C = \frac{1}{n-1}(X'^T X'), \tag{10}$$

where n is the number of data points (here, $n = 4$ for the four vertices). Performing eigenvalue decomposition $CV = \lambda V$ yields eigenvectors (principal directions) and eigenvalues (their variances). The eigenvector V_{max} associated with the largest eigenvalue λ_{max} defines the primary axis of the quadrilateral. Its angle relative to the x-axis is:

$$\theta = \arctan\left(\frac{V_{\max,y}}{V_{\max,x}}\right), \tag{11}$$

where $V_{\max,x}$ and $V_{\max,y}$ are the components of the eigenvector V_{max} in the x and y directions, respectively.

Based on the principal direction angle θ obtained via Principal Component Analysis (PCA), the vertices are rotated to align with the rectangular coordinate system. The dimensions of the rotated bounding box are calculated by projecting the rotated vertices (x'_i, y'_i) onto the principal direction and its orthogonal axis:

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x'_i \\ y'_i \end{bmatrix}, \tag{12}$$

where the width w and height h are defined as the differences between the maximum and minimum projected coordinates (u_i, v_i).

Using w, h, and θ, the precise parameters of the rotated bounding box are derived. Its center (x'_c, y'_c) is determined from the projected coordinates, and the corner coordinates (x''_i, y''_i) are obtained through rotation transformation:

$$\begin{bmatrix} x''_i \\ y''_i \end{bmatrix} = \begin{bmatrix} x'_c \\ y'_c \end{bmatrix} + \begin{bmatrix} \pm\frac{w}{2} \\ \pm\frac{h}{2} \end{bmatrix} \cdot \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{13}$$

where $\pm\frac{w}{2}$ and $\pm\frac{h}{2}$ represent the signed offsets of the corners relative to the center.

PCA determines the principal direction of maximum variance through eigenvalue decomposition of the covariance matrix. The eigenvector V_{max} corresponding to the largest eigenvalue defines this direction, providing a statistically optimal and robust estimate of the object's orientation that is invariant to noise and vertex ordering (Fig. 2).

IV. Experiment

A. Datasets

Our experiments are performed on three standard benchmarks for rotated object detection: DOTA-v1.0, HRSC2016, and DIOR-R.

DOTA-v1.0: This dataset is a large collection of optical remote sensing images (2806 in total), with dimensions spanning 800×800 to 4000×4000 pixels. It includes 188,282 annotated objects across 15 classes (e.g., PL, BD, BR, GTF, SV, LV, SH, TC, BC, ST, SBF, RA, HA, SP, HC).

HRSC2016 [6]: A specialized dataset for ship detection, HRSC2016 contains 1680 images with resolutions between 339×339 and 1333×1333 pixels.

DIOR-R [8]: Based on DIOR, this dataset focuses on oriented bounding boxes. It contains 23,463 images (800×800 pixels) with 192,518 instances across 20 categories.

TABLE II: Experimental results on HRSC2016. DQA* denotes DQA with PCA, implemented within the Oriented R-CNN framework using ResNet-101 backbone. mAP (07/12): VOC 2007/2012 metrics.

Method	**Model**	**mAP (07)**	**mAP (12)**
DRN	R101	-	92.70
Rol Trans	R101	86.20	-
AOPG	R101	90.34	96.22
R^{3}Det	R101	89.26	96.01
S^{2}ANet	R101	90.17	95.01
ReDet	ReR50	90.46	97.63
O-RCNN	R101	90.50	97.60
LSKNet	-	90.65	98.46
PKINet-S	-	90.70	98.54
STD-O	ViT-B	90.67	98.55
DQA*(ours)	R101	**90.80**	**98.90**

B. Main Results

Results on DOTA-v1.0: As shown in Table I, we comprehensively compare our method with more than 10 state-of-the-

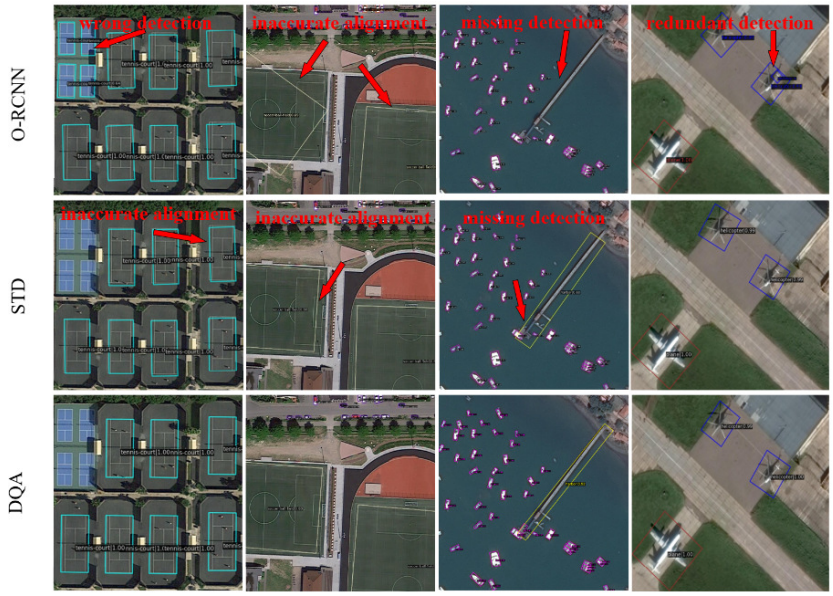

Fig. 3: Visual Results on DOTA-v1.0 Dataset. Top: O-RCNN; Middle: STD [5]; Bottom: Our DQA.

TABLE III: Experimental results on DIOR-R. DQA* denotes DQA with PCA, implemented within the Oriented R-CNN framework using ResNet-101 backbone.

Method	RetinaNet-O	RT [24]	LSKNet-S	GGHL
mAP	57.55	63.87	65.90	66.48
Method	Oriented Rep	DCFL	PKINet-S	**DQA***
mAP	66.71	66.80	67.03	**67.24**

art approaches on DOTA-v1.0. We embed DQA into the Oriented RCNN framework with the high-performance LSKNet as the backbone to achieve competitive results. Notably, DQA improves the AP by 12.56% for objects in the HC category, verifying the effectiveness of our detection head. Finally, DQA achieves a new state-of-the-art mAP of 79.38%.

Results on HRSC2016: We evaluated the performance of DQA on the HRSC2016 dataset using 10 of the most representative and prioritized methods. As shown in Table II, our method outperformed all other approaches, achieving mAPs of 90.80% and 98.90% according to the VOC 2007 and VOC 2012 metrics, respectively.

Results on DIOR-R: We provide a comparison of our DQA on the DIOR-R dataset, as shown in Table III. Our method achieves the best performance with an mAP of 67.24%.

Visualization Analysis: As shown in Fig. 3, the visual comparison on the DOTA-v1.0 dataset demonstrates that our DQA model achieves superior performance over O-RCNN and STD in terms of detection accuracy, bounding box alignment, and small object detection.

C. Ablation Study

To gain deeper insights into the performance of DQA and PCA, and their interactions, we conducted a series of ablation studies on DOTA-v1.0 using the O-RCNN as the detector. All backbones in this section were pre-trained on ImageNet-1K for 100 epochs to enhance performance.

TABLE IV: Ablation studies on DOTA-v1.0. DQA* denotes DQA with PCA.

(a) Detection head

BBoxHead	**mAP**	**FPS**	**Params (M)**	**FLOPs (G)**
2FC	75.87	**15.1**	**41.14**	**211.4**
DQA	76.38	14.5	42.68	216.3
DQA*	**77.84**	14.4	42.68	216.5

(b) Generalization across detectors

Detector	**Model**	**BBoxHead**	**mAP**
Rotated Faster R-CNN	ResNet-50	2FC	73.17
		DQA*	**75.06**
O-RCNN	ResNet-50	2FC	75.87
		DQA*	**77.84**

DQA Module: We compared the performance of three different detection head configurations, As shown in Table IVa. When DQA was introduced alone, the mAP increased by 0.51% (75.87% $\rightarrow$ 76.38%), indicating that the DQA attention mechanism effectively captures the geometric context of the objects.

Better performance often comes with mild extra computation, and our method maintains a favorable balance between accuracy and efficiency.This gain comes at a modest computational cost: +1.54 M parameters, +5.1 G FLOPs, and -0.7 FPS relative to the baseline. The resulting accuracy improvement justifies this trade-off, especially for challenging rotated objects.

While DQA effectively handles rotated objects, it suffers from limitations in cases of extreme aspect ratios or heavy occlusion, such as over-stretched quadrilaterals and confused

attention maps. We therefore introduce the PCA module to stabilize principal direction estimation and alleviate such geometric ambiguities.

Combined with PCA, the mAP is significantly boosted to 77.84% (+1.97%), showing that PCA effectively improves the robustness of principal direction prediction for quadrilaterals.DQA optimizes feature alignment while PCA resolves orientation ambiguity, and their collaboration substantially enhances detection accuracy.

DQA Adaptation to Different Backbone Architectures: Since DQA models the target's geometric properties through dynamic quadrilateral attention, our method is expected to be applicable to other RoI extraction methods. As shown in Table IVb, integrating the DQA module into the Rotated-RCNN object detector also demonstrates strong performance, highlighting the significant generalizability of our method.

V. Conclusion

In this paper, we propose a rotation object detection method based on the DQA mechanism. The DQA module dynamically adjusts the position, size, and orientation of the attention region through a data-driven window design, better adapting to the geometric properties of rotated objects. Additionally, the PCA-based main direction optimization module further improves the robustness of angle prediction and effectively addresses the detection challenges of irregular quadrilateral objects. Experimental results show that our method achieves significant performance improvements on the HRSC2016, DOTA-v1.0, and DIOR-R datasets, particularly for detecting direction-sensitive categories such as helicopters (HC), with a notable AP increase of 12.56%.

In future work, we will further explore the application of DQA in more scenarios (e.g., drone images, medical imaging) and optimize the model's computational efficiency to support real-time detection tasks.

Acknowledgments

This work was supported by the Open Fund of the Engineering Laboratory of Spatial Information Technology for Highway Geological Disaster Early Warning in Hunan Province (Grant No. kfj220803); the Natural Science Foundation of Hunan Province (Grants 2025JJ50394 and 2023JJ30050); the National Natural Science Foundation of China (Grants 61901061 and 62272062).

References

[1] Cai, X., Lai, Q., Wang, Y., Wang, W., Sun, Z., Yao, Y.: Poly kernel inception network for remote sensing detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27706–27716 (2024)

[2] Li, Y., Hou, Q., Zheng, Z., Cheng, M.-M., Yang, J., Li, X.: Large selective kernel network for remote sensing object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16794–16805 (2023)

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

[4] Han, J., Ding, J., Li, J., Xia, G.-S.: Align deep features for oriented object detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2021)

[5] Yu, H., Tian, Y., Ye, Q., Liu, Y.: Spatial transform decoupling for oriented object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence **38**(7), 6782–6790 (2024)

[6] Liu, Z., Wang, H., Weng, L., Yang, Y.: Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geoscience and Remote Sensing Letters **13**(8), 1074–1078 (2016)

[7] Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974–3983 (2018)

[8] Cheng, G., Wang, J., Li, K., Xie, X., Lang, C., Yao, Y., Han, J.: Anchor-free oriented proposal generator for object detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2022)

[9] Azimi, S.M., Vig, E., Bahmanyar, R., Körner, M., Reinartz, P.: Towards multi-class object detection in unconstrained remote sensing imagery. In: Asian Conference on Computer Vision, pp. 150–165. Springer (2018)

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

[11] Qian, W., Yang, X., Peng, S., Yan, J., Guo, Y.: Learning modulated loss for rotated object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence **35**(3), 2458–2466 (2021)

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

[13] Pu, Y., Wang, Y., Xia, Z., Han, Y., Wang, Y., Gan, W., Wang, Z., Song, S., Huang, G.: Adaptive rotated convolution for rotated object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6589–6600 (2023)

[14] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[15] Wang, D., Zhang, Q., Xu, Y., Zhang, J., Du, B., Tao, D., Zhang, L.: Advancing plain vision transformer toward remote sensing foundation model. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2022)

[16] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

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

[18] Zhang, Q., Zhang, J., Xu, Y., Tao, D.: Vision Transformer with Quadrangle Attention. arXiv preprint arXiv:2303.15105 (2023).

[19] Yang, S., Pei, Z., Zhou, F., Wang, G.: Rotated faster R-CNN for oriented object detection in aerial images. In: Proceedings of the 2020 3rd International Conference on Robot Systems and Applications, pp. 35–39 (2020)

[20] Yang, X., Yan, J., Feng, Z., He, T.: R3det: Refined single-stage detector with feature refinement for rotating object. In: Proceedings of the AAAI Conference on Artificial Intelligence **35**(4), 3163–3171 (2021)

[21] Dai, L., Liu, H., Tang, H., Wu, Z., Song, P.: Ao2-detr: Arbitrary-oriented object detection transformer. IEEE Transactions on Circuits and Systems for Video Technology **33**(5), 2342–2356 (2022)

[22] Zeng, Y., Yang, X., Li, Q., Chen, Y., Yan, J.: Ars-detr: Aspect ratio sensitive oriented object detection with transformer. arXiv preprint arXiv:2303.04989 (2023)

[23] Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. arXiv preprint arXiv:1904.01355 (2019)

[24] Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M.Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., et al.: Object detection in aerial images: A large-scale benchmark and challenges. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7778–7796 (2021)