FRED: Towards a Full Rotation-Equivariance in Aerial Image Object Detection

Chanho Lee[1], Jinsu Son[1], Hyounguk Shon[1], Yunho Jeon[2], Junmo Kim[1]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]Hanbat National University, South Korea
{yiwan99, sonjs, hyounguk.shon}@kaist.ac.kr, yhjeon@hanbat.ac.kr, junmo.kim@kaist.ac.kr

Abstract

Rotation-equivariance is an essential yet challenging property in oriented object detection. While general object detectors naturally leverage robustness to spatial shifts due to the translation-equivariance of the conventional CNNs, achieving rotation-equivariance remains an elusive goal. Current detectors deploy various alignment techniques to derive rotation-invariant features, but still rely on high capacity models and heavy data augmentation with all possible rotations. In this paper, we introduce a Fully Rotation-Equivariant Oriented Object Detector (FRED), whose entire process from the image to the bounding box prediction is strictly equivariant. Specifically, we decouple the invariant task (object classification) and the equivariant task (object localization) to achieve end-to-end equivariance. We represent the bounding box as a set of rotation-equivariant vectors to implement rotation-equivariant localization. Moreover, we utilized these rotation-equivariant vectors as offsets in the deformable convolution, thereby enhancing the existing advantages of spatial adaptation. Leveraging full rotation-equivariance, our FRED demonstrates higher robustness to image-level rotation compared to existing methods. Furthermore, we show that FRED is one step closer to non-axis aligned learning through our experiments. Compared to state-of-the-art methods, our proposed method delivers comparable performance on DOTA-v1.0 and outperforms by 1.5 mAP on DOTA-v1.5, all while significantly reducing the model parameters to 16%.

Introduction

Aerial object detection is an emerging field in the domain of computer vision. Since aerial images capture objects with arbitrary orientations and are often densely packed, oriented bounding box (OBB) can provide a tighter representation in such cases. One distinguishing feature of aerial images is the non-axis aligned nature, absence of any top-bottom or left-right bias. The ideal aerial object detector should consistently deliver predictions irrespective of object orientation. If the image undergoes rotation, the predicted OBB should also rotate concurrently. Hence, for an oriented object detector to be reliable, it must exhibit rotation-equivariance.

However, achieving rotation-equivariance on oriented object detection is challenging, since most researches are extended from horizontal object detection models. Most methods rely on the assumption of accurate orientation estimation and focus on making features invariant to rotation. A representative approach to achieve this is the RoI Transformer (Ding et al. 2019), which leverages rotation-sensitive region-of-interest (RoI) pooling on a rotated RoI (RRoI) to acquire instance-level rotation-invariance. Such orientation-specific feature refinement has demonstrated its efficiency across one-stage object detectors (Han et al. 2021a) and anchor-free detectors (Pan et al. 2020). Another strategy is using point set representation which implicitly represents the oriented bounding box as a set of adaptively learned points (Guo et al. 2021; Li et al. 2022). These approaches have managed to separate out the prediction of orientation itself, naturally leads to non-axis aligned feature learning. Yet, these aforementioned methods rely heavily on data augmentation using random rotations and remain distant from achieving true rotation-equivariance.

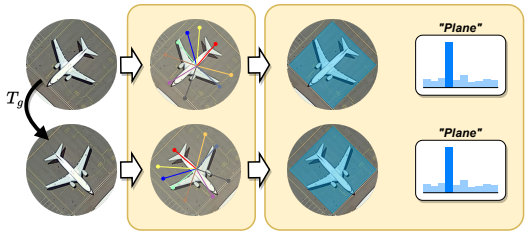


Figure 1: Overview of the fully rotation-equivariant object detector (FRED). FRED consists of a rotation-equivariant backbone which predicts a point set followed by two prediction branches – (1) a rotation-equivariant box regression head and (2) a rotation-invariant classification head. We achieve end-to-end equivariance for object detection.

Recently, Han et al. (2021b) proposed ReDet, a rotation-equivariant detector firstly employing rotation-equivariant CNNs (Weiler and Cesa 2019). Their Rotation-invariant ROI Align (RiRoI Align) leverages the characteristics of rotation-equivariant features, enabling the extraction of rotation-invariant features dependent on the rotated RoI. However, it is worth noting that even though their RiRoI Align operates based on a rotation-equivariant theory, the orientation of the predicted RRoI itself is not equivariant. Due to the ambiguity and angular discontinuity of OBB rep-

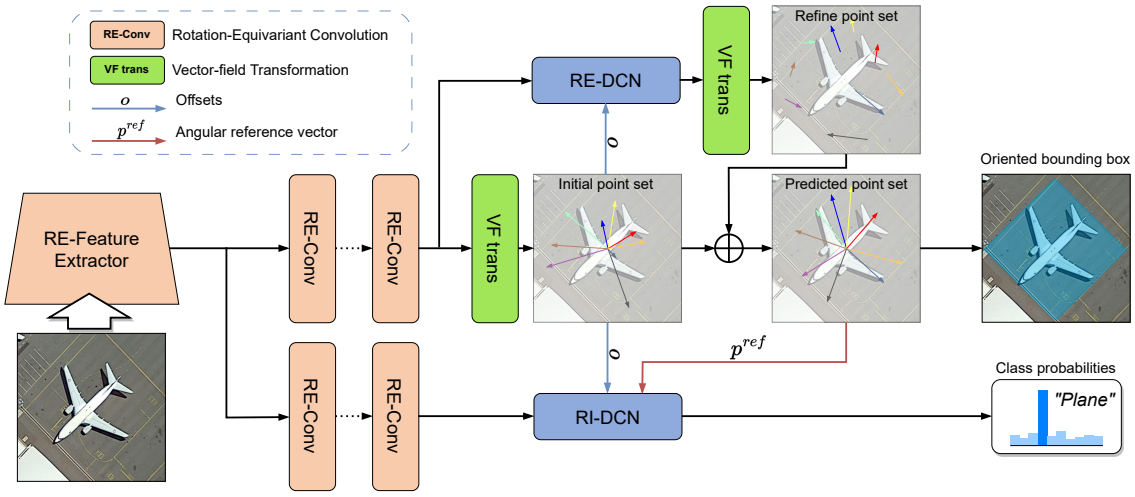


Figure 2: Overall model architecture of the proposed Fully Rotation-Equivariant Detector (FRED). C_N-equivariant features are fed into the rotation-equivariant head up to two deformable convolution blocks. The Rotation-Equivariant Deformable Convolution (RE-DCN) tilizes an initial point set as an offset and refines it through spatial adaptation without breaking rotation-equivariance. The Rotation-Invariant Deformable Convolution (RI-DCN) performs an orientation alignment to produce rotation-invariant features using an align reference vector sourced from the localization branch. As both the deformable offsets and the reference vector maintain rotation-equivariance, the classification branch achieves instance-level rotation-invariance.

resentations, rotation-invariance of RiRoI Align is vulnerable to significant degrees of rotations. To design a model that remains consistent across any rotation, we utilize a point set representation in place of the bounding box, and endow the entire localization process with strict rotation-equivariance.

In this work, we propose a fully rotation-equivariant oriented object detector named *FRED* which leverages point set representation to achieve full rotation-equivariance on both classification and localization. We conceptualize the bounding box as a set of rotation-equivariant vectors. By employing this idea, we ensure that with any rotation, the vectors not only shift in accordance with the image-level rotation but also change their orientation simultaneously. This trait perfectly satisfies the attributes needed for oriented bounding box prediction as depicted in Figure 1. Furthermore, we apply these rotation-equivariant vectors as offsets of deformable convolution. This allows us to propose Rotation-Equivariant Deformable Convolution (RE-DCN) and Rotation-Invariant DCN (RI-DCN), which can simultaneously achieve spatial and orientation alignment through an rotation-equivariant receptive field. Compared to previous methods, our FRED is highly robust to image rotations powered by end-to-end rotation-equivariance. Moreover, FRED maximizes the benefits of the high-level weight sharing of rotation-equivariant CNNs, showcasing superior performance with fewer learnable parameters than any other detectors.

Moreover, we discovered a promising phenomenon while training our rotation-equivariant model. Typically, rotation-equivariance is associated with robustness to different rotations for a single instance. If there exists an instance group with similar context and scale, we can anticipate rotation-equivariance among them. We observed that FRED, just before full convergence, learned the relative pose of objects without any direction-specific supervision. While this tendency diminishes during the convergence process, it can be seen as a reflection of FRED being trained in a genuinely non-axis aligned manner. Through our experiments, we demonstrate that previous non-axis aligned methods are still being trained in an axis-overfitted manner, while our FRED showcases a more genuine non-axis aligned learning.

In summary, our main contributions in this paper are as follows:

- To the best of our knowledge, we are the first to propose a fully rotation-equivariant oriented object detector. Compared to previous state-of-the-art methods, FRED guarantees more robust predictions against image rotations.
- We propose novel methods that combine deformable convolution and rotation-equivariant vectors to simultaneously perform spatial and orientation alignment without disrupting equivariance.
- Our experiments demonstrate that FRED achieves promising results with significantly fewer parameters, and offers a new insight into axis-free learning.

Related Works

Oriented Object Detection

The main approach for oriented object detection extends from horizontal object detection with additional regression for orientation. Challenges in oriented object detection in aerial images arise from arbitrary oriented and densely packed objects. RoI Transformer (Ding et al. 2019) proposed a rotation-sensitive RoI pooling for obtaining rotation-invariance. SCRDet (Yang et al. 2019a), DRN (Pan et al. 2020), S2A-Net (Han et al. 2021a), and R^3Det (Yang et al.

2021) adressessed the challenges through methods that refine features. Most research focuses on developing methods to apply the axis-aligned property to non-axis aligned oriented object detection, yet rely on well predefined anchors with angular discontinuity of OBB.(Qian et al. 2021; Yang et al. 2022)

Alternative Bounding Box Representations

The introduction of point set representation offers a promising alternative to these issues, escaping from traditional anchors and bounding boxes. By adaptively capturing object context, point set has demonstrated their capability to learn richer representations as shown in RepPoints (Yang et al. 2019b). Through strategies like convex-hull feature adaptation (Guo et al. 2021) and orientation-sensitive sampling (Li et al. 2022), point set based methods outperform the previous anchor-based detectors. Our FRED introduces rotation-equivariant point set prediction and its benefit, showcasing its closer alignment with true non-axis aligned learning.

Rotation-Equivariant Neural Networks

Beginning with the Group-Equivariant CNN proposed by Cohen and Welling (2016), several methods have introduced rotation-equivariant CNNs using steerable filters (Weiler and Cesa 2019; Cesa, Lang, and Weiler 2021) and they have been proven effective in various imagery fields (Veeling et al. 2018; Gupta, Arya, and Gavves 2021; Lee et al. 2023). Recently, Han et al. (2021b) introduced ReDet, a rotation-equivariant detector for oriented object detection that firstly utilizes the rotation-equivariant CNNs. While RiRoI Align offers a rotation-invariant transform, its dependency on the non-equivariantly predicted RRoI still introduces a residual challenge to achieving full rotation-equivariance.

Preliminaries

This section offers a brief overview on the concept of rotation-equivariance. Given the prevalent use of steerable filters to yield rotation-equivariant features, the finite rotation group and its representation through group-wise permutation are introduced.

Let G be a group which can be any transformations on image space X. Then a function $\Phi : X \to Y$ is said to be *equivariant* if

$$\Phi\left(T_g^X(x)\right) = T_g^Y\left(\Phi(x)\right) \quad \forall g \in G, \forall x \in X \tag{1}$$

where T_g^X and T_g^Y is a group action defined on each space. If T_g^Y is identity mapping, then invariance holds. For example, the conventional CNN Φ shares convolution weight at every location, so satisfies the equation above on translation action.

Rotation-Equivariance

In this paper, we are addressing rotation-equivariance, so group can be formulated as the semi-direct product of the translation group $(R^2, +)$ and rotation group H. For a group G, the rotation-equivariance of function $f : X \to Y$ can be expressed as

$$f\left(T_g^X(x)\right) = T_g^Y\left(f(x)\right) \quad \forall g \in G, \forall x \in X \tag{2}$$

given T_g^X and T_g^Y as rotation action on X and Y respectively. If T_g^X and T_g^Y are isomorphic image-level rotations in each space, then function f can be viewed as transforming the image space into a rotation-equivariant scalar field.

On the other hand, let a function $v : X \to V$ be a mapping from the image space X to 2-dimensional vector field space V. To provide clearness in notation, let us denote T_g as the group action of G that operates on both X and V. Then rotation-equivariance of vector field can be expressed as:

$$v\left(T_g(x)\right) = R_g T_g\left(v(x)\right) \quad \forall g \in G, \forall x \in X \tag{3}$$

In this context, R_g is a group action on G that rotates every vector in parallel with T_g. For clarity, a rotation-equivariant vector field necessitates not just image-level rotations T_g but also ensures that the vector predicted at each image pixel rotates by R_g.

Steerable Filters and Cyclic-Equivariance

A commonly used method to implement a practical rotation-equivariant CNN is to utilize steerable filters. If a steerable filter rotates at intervals of $2\pi/N$ degree and forms N weight-shared filters, we can create a convolution layer that is discretely rotation-equivariant to the cyclic group C_N. For a group action T_g on the cyclic group C_N, cyclic-equivariance can be expressed as:

$$f\left(T_g(x)\right) = P_g T_g\left(f(x)\right) \quad \forall g \in C_N, \forall x \in X \tag{4}$$

where P_g is rotation group-wise permutation operator. For example, when the image is rotated by $2\pi n/N$, the C_N-equivariant feature undergoes both an image-level rotation and a group-wise permutation of degree n. It's worth noting that every intermediate feature of a rotation-equivariant CNN is cyclic-equivariant. One major property of cyclic-equivariant feature is that it can be transformed into either rotation-invariant or rotation-sensitive features through a pooling operation. As in Weiler and Cesa (2019), rotation-invariant features are obtained from the maximum response across all rotations, achieved through rotation group-wise max pooling such as $\max_g f(x)$.

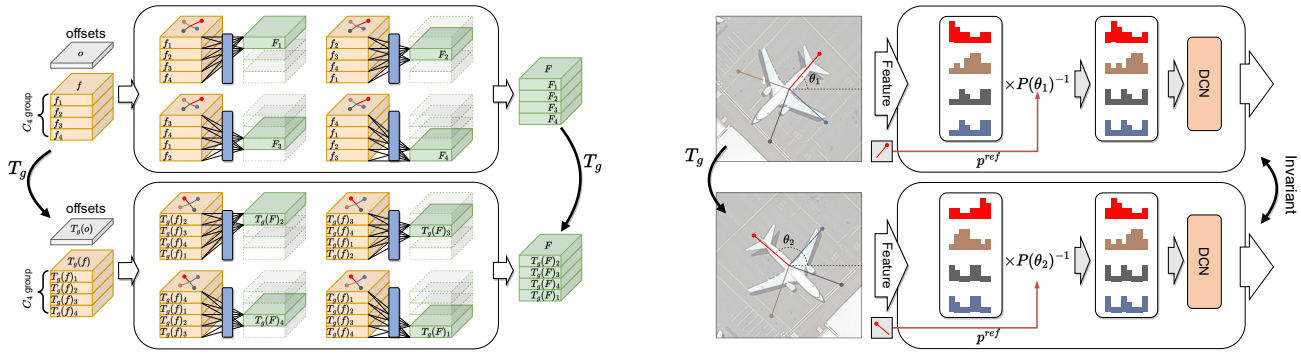
On the other hand, rotation-equivariant vectors can be obtained not just from the max response, but also using the argmax operator. By combining max response $\max_g f(x)$ and its orientation $\theta = \frac{2\pi}{N} \arg\max_g f(x)$, we can transform a cyclic-equivariant feature to a vector field $v(x)$ as

$$v(x) = \left(\max_g f(x)\right) \cdot [\cos(\theta), \sin(\theta)]^T \tag{5}$$

For distinguishing the rotation-equivariant vector field, we will call the rotation-equivariant scalar field as instance-level rotation-invariance. For a comprehensive overview of group-equivariant CNNs, we refer readers to Weiler and Cesa (2019).

Methodology

Initially, rotation-equivariant features are extracted from the rotation-equivariant backbone and neck. To ensure



(a) Rotation-equivariant Deformable Convolution

(b) Rotation-invariant Deformable Convolution

Figure 3: This example illustrates a 4-equivariant rotation group (C_4) and 2x2 deformable kernel for simplicity. The deformable convolution (DCN) layer parameters are shared between rotation groups.

rotation-equivariant localization with the point set $S = \{[x^k, y^k]^\top\}_{k=1}^K$, we transform cyclic-equivariant features into K vector fields as Equation (5), through a vector-field transformation layer. Our core idea is to utilize this point set as an offset for deformable convolution, enabling a rotation-equivariant adaptive receptive field. In the localization branch, Rotation-Equivariant Deformable Convolution (RE-DCN) maintains cyclic-equivariance to perform localization refinement through spatial adaptation. On the other hand, the Rotation-Invariant Deformable Convolution (RI-DCN) transforms features into invariant ones to ensure robust classification. We have advanced the orientation alignment proposed in ReDet (Han et al. 2021b), obtaining better rotation-equivariance and strengthening the connection between localization and classification through our alignment reference vector driven by point set prediction. Finally, we introduce the edge constraint loss to assure orientation sensitivity and stable training of the alignment reference vector. The overall architecture of our proposed method is depicted in Figure 2.

Rotation-Equivariant Deformable Convolution

The key role of rotation-equivariant offset is to guarantee that each kernel of the deformable convolution consistently focuses on a semantically identical area. However, directly applying regular deformable convolution breaks rotation-equivariance since it blends the features of all rotation groups, being agnostic to their distinctions. To solve this, we introduce Rotation-Equivariant Deformable Convolution (RE-DCN) which allows all N rotation groups to perform independently. As shown in Figure 3a, each rotation group independently executes deformable convolution using shared offsets and weights. Through the use of RE-DCN, the output feature retains cyclic-equivariance, allowing it to be transformed into a rotation-equivariant vector field. This approach not only ensures spatial feature refinement without compromising rotation-sensitivity but also achieves an N-factor parameter efficiency through weight sharing. A detailed structure and the corresponding pseudo code are provided in the appendix.

Rotation-Invariant Deformable Convolution

The classification branch needs consistent prediction that comes from rotation-invariance, rather than retaining the feature's orientation. At a basic level, if we make the feature that is input to the deformable convolution rotation-invariant, the output can also achieve invariance. A potential approach to this is the rotation group-wise max pooling mentioned earlier (Cohen and Welling 2016). However, this strategy leads to a significant reduction in the model's capacity due to a dramatic channel reduction, and it also sweeps away all sensitive orientation information.

To tackle this issue, ReDet (Han et al. 2021b) introduced the orientation alignment, which can be formulated as:

$$OA(f, \theta) = Int\left(SC(f, r), \theta\right), \quad r = \lfloor \theta N / 2\pi \rfloor \tag{6}$$

where $Int(\cdot)$ denotes bilinear interpolation between adjacent rotation groups and $SC(\cdot)$ is rotation group-wise permutation. In simpler terms, orientation alignment can be seen as an inverse direction of cyclic permutation P_g, aiming to achieve rotation-invariance. While the orientation alignment is itself a rotation-equivariant operator, it is the θ that truly determines rotation-invariance. It is important to highlight that the orientation alignment introduced in ReDet has certain shortcomings in ensuring rotation-invariance. Specifically, (a) the θ is derived from the RoI head, which comprises regular CNNs, so eventually loses its rotation-equivariance, and (b) even if θ is predicted with utmost accuracy, its efficacy still hinges on the specific representation style of the oriented bounding box and predefined anchors.

In contrast, each individual point of our predicted localization is rotation-equivariant, making it suitable for orientation alignment. Our approach ensures that (a) through our proposed localization branch, θ naturally attains rotation-equivariance, (b) the network can adaptively learn the appropriate direction without constraints imposed by the OBB representation style, and (c) it strengthens the correlation between localization and classification. To ensure robustness to the feature grid location, we define the angular reference vector $\vec{p}^{ref}$ for orientation alignment as $\vec{p}^{ref} = [x^{ref} - x^c, y^{ref} - y^c]^\top$, where $[x^{ref}, y^{ref}]^\top$ is the alignment

reference point and $[x^c, y^c]^\top$ is the center point of the convex hull formed through the point set $S = \{[x^k, y^k]^\top\}_{k=1}^K$, respectively. Then, our oriented alignment is defined as $OA(f, \measuredangle \vec{p}^{ref})$ where $\measuredangle(\cdot)$ is the angle from the x-axis.

We emphasize the fact that our point set prediction is made up of rotation-equivariant vectors. This means any point within the set can serve as an alignment reference, ensuring rotation-invariance of alignment. Through our proposed orientation alignment, the transformed rotation-invariant feature is fed into the deformable convolution and utilized for classification. We refer to this entire process as Rotation-Invariant Deformable Convolution (RI-DCN), which is depicted in Figure 3b.

Orientation Alignment without Degeneration

Although every rotation-equivariant point can be utilized as an alignment reference, we found that randomly selecting a point can sometimes lead to significant training instability. In point set based methods (Yang et al. 2019b; Guo et al. 2021; Li et al. 2022), we observed that a few points might be always excluded from the convex-hull, and wander around the center of the object. Although such points still remain rotation-equivariant, their direction can change dramatically during training, potentially leading to noisy alignment and unstable learning. Therefore, without majorly impacting the existing localization performance, there is a need to ensure that the reference point is influential in both localization and the formation of the convex hull. We introduce the edge constraint loss, pushing the reference point towards the nearest center points of the four edges of the ground truth bounding box. Given the four center points from the ground truth box edges, denoted as $\{[x_i^g, y_i^g]^\top\}_{i=1}^4$, the edge constraint loss $\mathcal{L}_{ec}$ can be formulated as

$$\mathcal{L}_{ec} = \min_i \left\| [x^{ref}, y^{ref}]^\top - [x_i^g, y_i^g]^\top \right\|_2 \tag{7}$$

We set a default weight of edge constraint loss as 0.025 to minimize its influence on localization loss. Without losing generality, we can set the first point of the predicted point set as the alignment reference.

Experiments

Benchmark and Implementation Details

DOTA dataset (Xia et al. 2018; Ding et al. 2021) is a large-scale benchmark designed for assessing oriented object detection in aerial images. It includes 2,806 aerial images with sizes varying from 800×800 to 4000×4000. In **DOTA-v1.0**, there are 188,282 instances distributed among 15 specific categories: Plane (PL), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), 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). **DOTA-v1.5** shares the image set with DOTA-v1.0 but introduces an additional class: Container Crane (CC), bringing the total to 402,089 instances. With the inclusion of extremely small objects, DOTA-v1.5 is more challenging, yet it seems to provide more reliable labels. For experimental settings, both the training and validation sets from DOTA are combined for training, with the test set reserved for evaluations. Images are typically split into 1024×1024 patches using an 824 stride. For evaluation metric, the mean average precision (mAP) metric (Everingham et al. 2010) is used.

Model Architecture. Our implementation is based on the MMRotate (Zhou et al. 2022) and $E(2)$-CNN (Weiler and Cesa 2019) framework. Our proposed model is based on C_8 ReResNet-50 backbone pretrained on ImageNet, with ReFPN neck proposed in Han et al. (2021b). We stacked three C8 rotation-equivariant convolution layers for each branch, and used modulated deformable convolution (Zhu et al. 2019) for the classification branch. Focal loss (Lin et al. 2017), convex IOU loss (Rezatofighi et al. 2019) and spatial constraint loss with APAA strategy as described in Li et al. (2022) is employed for training. We set the weight of our edge constraint loss as 0.0025 as default.

Training Scheme. The training was conducted with the stochastic gradient descent optimizer with the momentum and the weight decay set to 0.9 and 0.0001, respectively. The initial learning rate is 0.008, and the model is trained for 40 epochs with batch size 8, using a step decay schedule.To ensure a fair comparison, we refrained from fine-tuning any hyper-parameters related to losses, sampling strategies, and training schedules, in line with the settings from Guo et al. (2021) and Li et al. (2022).

Strided Convolution and Equivariance. Due to the discreteness of images, rotation-equivariant CNNs are strictly equivariant only for rotations that are multiples of 90 degree. However, employing even-sized images and strided convolution layers can disrupt this strict equivariance, even for 90-degree rotations. This issue arises because strided convolution always samples the top-left corner, as reported by Romero et al. (2020). While the impact on performance might be minimal, given our goal to achieve rotation robustness through not approximate but perfect rotation-equivariance, we decided to pursue a more stringent rotation-equivariance. To address this issue, by simply adding zero-padding to ensure strided convolution always experience odd-numbered images, strict equivariance can be achieved without any significant modifications to the network structure. A more detailed analysis of how strided convolution can disrupt rotation-equivariance is covered in the appendix.

Comparison with the State-of-the-art Methods

DOTA-v1.0. Based on the experimental results on the single-scale DOTA-v1.0 dataset, we aim for a fair comparison with previous methods. Our model, using the ReResNet50 and ReFPN backbone, shows comparable result ranked second among anchor-free methods. we note that our model consists of only 16% of model size.

DOTA-v1.5. Similar to DOTA-v1.0, we also report single-scale experiments for DOTA-v1.5. For a fair comparison, we reimplemented OrientedRepPoints using the official code. Our model achieved 78.3 mAP and surpasses the state-of-the-art anchor-free method by 1.4 mAP. Given that our model demonstrates superior results in more difficult and reliable dataset settings, we assert that it possesses a greater

Methods	Model	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Anchor-based:																	
RoI Trans.	R101	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	69.56
FAOD	R101	**90.21**	79.58	45.49	76.41	73.18	68.27	79.56	90.83	83.40	84.68	53.40	65.42	74.17	69.69	64.86	73.28
SCRDet	R101	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**	72.61
Gliding Vertex	R101	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	75.02
R^3Det	R101	88.76	83.09	50.91	67.27	76.23	80.39	86.72	90.78	84.68	83.24	61.98	61.35	66.91	70.63	53.94	73.79
S^2A-Net	R50	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
DAL	R101	88.61	79.69	46.27	70.37	65.89	76.10	78.53	90.84	79.98	78.41	58.71	62.02	69.23	71.32	60.65	71.78
KFIoU(R^3Det)	R50	89.05	75.17	49.04	69.67	78.06	75.46	86.69	**90.90**	83.65	84.48	62.21	62.87	66.72	65.95	50.20	72.68
ReDet	ReR50	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**
Anchor-free:																	
PIoU	DLA34	80.90	69.70	24.10	60.20	38.30	64.40	64.80	**90.90**	77.20	70.40	46.50	37.10	57.10	61.90	64.00	60.50
O^2-DNet	H104	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	71.04
DRN*	H104	**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	73.23
CFA	R101	89.26	81.72	51.81	67.17	79.99	78.25	84.46	90.77	83.40	85.54	54.86	67.75	73.04	70.24	64.96	75.05
RSDet++	R152	86.80	82.70	**54.60**	71.70	76.00	71.20	83.50	87.40	83.40	85.30	**72.40**	62.90	70.90	72.00	**70.40**	75.40
OrientedReps	R50	87.02	**83.17**	54.13	71.16	**80.18**	**78.40**	87.28	**90.90**	85.97	**86.25**	59.90	**70.49**	**73.53**	**72.27**	58.97	**75.97**
FRED (Ours)	ReR50	89.37	82.12	50.84	**73.89**	77.58	77.38	**87.51**	90.82	**86.30**	84.25	62.54	65.10	72.65	69.55	63.41	75.56

Table 1: Comparisons with state-of-the-art methods on DOTA-v1.0 OBB Task. R50, R101, H104, and ReR50 mean ResNet50, ResNet101, Hourglass104 and rotation-equivariant ResNet50. The best and second-best results are boldfaced and underlined, respectively.

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP
RetinaNet-O	71.43	77.64	42.12	64.65	44.53	56.79	73.31	90.84	76.02	59.96	46.95	69.24	59.65	64.52	48.06	0.83	59.16
Mask R-CNN	76.84	73.51	49.90	57.80	51.31	71.34	79.75	90.46	74.21	66.07	46.21	70.61	63.07	64.46	57.81	9.42	62.67
HTC	77.80	73.67	51.40	63.99	51.54	73.31	80.31	90.48	75.12	67.34	48.51	70.63	64.84	64.48	55.87	5.15	63.40
RoI-Trans.	71.70	82.70	**53.00**	71.50	51.30	74.60	80.60	90.40	78.00	68.30	53.10	**73.40**	**73.90**	65.60	56.90	3.00	65.50
ReDet	79.20	**82.81**	51.92	71.41	52.38	**75.73**	80.92	90.83	75.81	68.64	49.29	72.03	73.36	**70.55**	**63.33**	11.53	66.86
OrientedReps	75.52	82.60	51.24	70.21	57.81	73.82	**86.25**	**90.86**	78.30	**76.47**	**53.61**	72.78	66.68	69.48	53.66	11.09	66.90
FRED (Ours)	**79.60**	81.44	52.60	**72.57**	**58.07**	74.82	86.12	90.81	**82.13**	74.84	53.37	72.93	69.51	69.91	54.82	**19.27**	**68.30**

Table 2: Comparisons with state-of-the-art methods on DOTA-v1.5 OBB Task. RetinaNet-O, Mask R-CNN, HTC, and ReDet reported from (Han et al. 2021b).

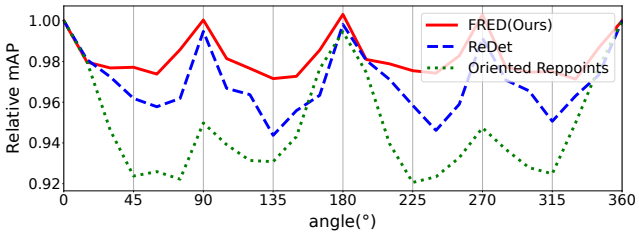


Figure 4: Robustness against rotation estimated with DOTA-v1.0. We compared the performance degradation of various models as the image rotates. Excluding discrete rotations at 90-degree intervals, the loss of rotated image information always results a decreased mAP.

	Method				
Vector-field	-	-	✓	✓	✓
GroupPool	-	✓	✓	-	-
OA	-	-	-	✓	✓
Edge constraint	-	-	-	-	✓
mAP	73.32	73.91	74.69	75.35	75.56

Table 3: Ablations incrementally added equivariances. "Vector-field" refers to vector-field transformation layer.

generalization capability.

Robustness against rotations. The goal of our fully rotation-equivariant model is to provide consistently reliable predictions during the inference phase regardless of rotations. We observed how the performance changes by varying the rotation of the test images. As seen in Figure 4, while our method consistently performs regardless of image rotation, non-equivariant models exhibit significant variations in performance. Specifically, while the previous methods show performance degradation for large rotations, our model comparably maintains its performance. Even though our model is designed to be discretely rotation-equivariant at 45-degree intervals, it exhibits approximate rotation-equivariance across every continuous rotations.

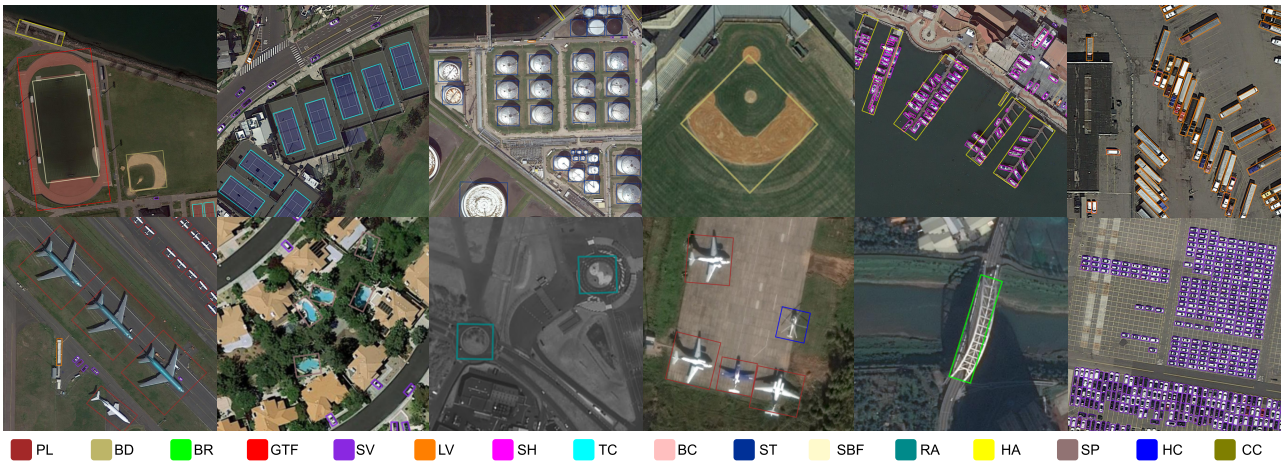
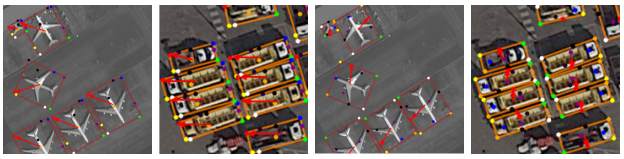


Figure 5: Examples of detection results using FRED on DOTA-v1.5



(a) Oriented RepPoints (b) FRED (Ours)

Figure 6: Directionality emerges with the equivariant point set representation. Compared to Oriented RepPoints (Li et al. 2022), FRED maintains object orientation without explicit supervision. For visualization purposes, we color-coded each point in the predicted point set with respect to the ordering. Arrow indicates the first point in the set. Both models are trained for 30 epochs and early-stopped before convergence.

Methods	Backbone	mAP	Size (MB)
Oriented RepPoints	Res50-FPN	66.90	140.78
ReDet	ReRes50-ReFPN	66.86	124.97
FRED (Ours)	ReRes50-ReFPN	**68.30**	**22.28**

Table 4: Comparison of model size on DOTA-v1.5

Discussion

Ablation on rotation-equivariances. To understand the effectiveness of rotation-equivariance for both object classification and localization, we conducted experiments on the DOTA-v1.0 dataset. To gauge the efficacy of rotation-equivariance on each branch, we measured performance by incrementally adding various rotation-equivariances based on the ReResNet50 and ReFPN feature extractor.

In Table 3, the term 'Vector-field' refers to the implementation of a rotation-equivariant vector-field in the localization branch. GroupPool and Orientation Alignment (OA) are both rotation-invariance for classification branch, but with different way. The efficiency of rotation-equivariance for both classification and localization is evident. Without any equivariance transformation, a model simply becomes in a capacity reduced model with rotation-agnostic behavior. Since group pooling removes every orientation sensitivity and thus exhibits relatively lower performance.

Parameter Efficiency. Our method leverages rotation-equivariant convolution across all layers, from the backbone to the head, to ensure full rotation-equivariance. This maximizes the benefits of weight sharing, allowing us to outperform existing models with only 16% of the parameters. As seen in Table 4, even when compared to ReDet, which utilizes the same rotation-equivariant backbone, FRED has only 18% of the size.

Observations on point set directionality. Originally, rotation-equivariance is defined for various rotations of a single instance and is not applied to different instances. If rotation-equivariance also operates between multiple objects of the same class, we expect FRED to understand their relative orientations without any supervision. We observed that FRED captures the relative pose between objects during training (see Figure 6), even though its performance is lower than fully trained model. This tendency tends to diminish during the training process and only weakly appears in the converged model. We hypothesize that rotation-equivariance might coarsely cluster objects with similar shapes, sizes, and colors in the early stages of training, naturally making the model aware of their poses. However, since objects within the same class can have varied distributions, those pose sensitivity derived from coarse rotation-equivariance fades away. Nevertheless, this distinctly indicates that FRED engages in non-axis aligned feature learning. In contrast, the previous point set based methods produce predictions solely based on the bounding box distribution, agnostic to the object's direction. Even if they are not aligned to the horizontal or vertical axis, the predictions appear to be aligned to a fixed axis inherent to each point. We may leave this intriguing behavior of FRED as our future work.

Conclusion

In this paper, we introduced a fully rotation-equivariant object detector for aerial images. Our novel rotation-equivariant deformable convolution blocks offer improved alignment for spatial region and orientation with rotation-equivariant receptive fields. Moreover, through rotation-invariant orientation alignment, we strengthened the correlation between classification and localization. Not only does our model display rotation robustness, parameter efficiency, and promising quantitative results, but it also hints at the potential for unsupervised pose estimation.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number : HI20C1234).

References

Cesa, G.; Lang, L.; and Weiler, M. 2021. A program to build E (N)-equivariant steerable CNNs. In *International Conference on Learning Representations*.

Cohen, T.; and Welling, M. 2016. Group equivariant convolutional networks. In *International conference on machine learning*, 2990–2999. PMLR.

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/CVF Conference on Computer Vision and Pattern Recognition*, 2849–2858.

Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2021. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Guo, Z.; Liu, C.; Zhang, X.; Jiao, J.; Ji, X.; and Ye, Q. 2021. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 8792–8801.

Gupta, D. K.; Arya, D.; and Gavves, E. 2021. Rotation equivariant siamese networks for tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12362–12371.

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

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

Lee, J.; Kim, B.; Kim, S.; and Cho, M. 2023. Learning Rotation-Equivariant Features for Visual Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21887–21897.

Li, W.; Chen, Y.; Hu, K.; and Zhu, J. 2022. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1829–1838.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; and Xu, C. 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.

Qian, W.; Yang, X.; Peng, S.; Yan, J.; and Guo, Y. 2021. Learning modulated loss for rotated object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2458–2466.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Romero, D.; Bekkers, E.; Tomczak, J.; and Hoogendoorn, M. 2020. Attentive group equivariant convolutional networks. In *International Conference on Machine Learning*, 8188–8199. PMLR.

Veeling, B. S.; Linmans, J.; Winkens, J.; Cohen, T.; and Welling, M. 2018. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, 210–218. Springer.

Weiler, M.; and Cesa, G. 2019. General e (2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, X.; Yan, J.; Feng, Z.; and He, T. 2021. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3163–3171.

Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; and Fu, K. 2019a. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8232–8241.

Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; and Tian, Q. 2022. The KFIoU loss for rotated object detection. *arXiv preprint arXiv:2201.12558*.

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019b. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9657–9666.

Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. 2022. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7331–7334.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9308–9316.