R^{3}Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object

Xue Yang[1], Junchi Yan[1,*], Ziming Feng[2], Tao He[3,4]

{yangxue-2019-sjtu, yanjunchi}@sjtu.edu.cn, zimingfzm@cmbchina.com, Tommie.he@cowarobot.com,

[1]Department of Computer Science and Engineering, and MoE Key Lab of Artificial Intelligence, AI Institute Shanghai Jiao Tong University

[2]China Merchants Bank Credit Card Center. [3]Anhui COWAROBOT CO., Ltd.

[4]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation

Introduction

- **Rotation Object Detection**

➢ Task: design a novel multi-category rotation detector for small, cluttered and rotated objects.

➢ Challenges

- **Large aspect ratio**. The Skew Intersection over Union (SkewIoU) score between large aspect ratio objects is sensitive to change in angle.
- **Densely arranged**. Many objects usually appear in densely arranged forms.
- **Arbitrary orientations**. Objects in images can appear in various orientations, which requires the detector to have accurate direction estimation capabilities.

- **Our main contributions**

➢ For large aspect ratio object detection, an accurate and fast rotation singe-stage detector is devised in a refined manner, for high-precision detection. In contrast to the recent learning based methods for feature alignment, which lacks an explicit mechanism to compensate the misalignment, we propose a direct and effective pure computing based approach which is further extended to handle the rotation case.

➢ For densely arranged objects, we develop an efficient coarse-to-fine progressive regression approach to better exploring the two forms of anchors in a more flexible manner, tailored to each detection stage.

➢ For arbitrarily-rotated objects, a derivable approximate SkewIoU loss is devised for more accurate rotation estimation.

➢ Codes: https://github.com/Thinklab-SJTU/R3Det_Tensorflow

Proposed Approach

- **Pipeline**

We give an overview of our method as sketched in Fig. 1. The embodiment is a refined single-stage rotation detector based on the RetinaNet, namely Refined Rotation RetinaNet (R^{3}Det). The refinement stage (which can be added and repeated by multiple times) is added to the network to refine the bounding box, and the feature refinement module FRM is added during the refinement stage to reconstruct the feature map. In a single-stage rotating object detection task, continuous refinement of the predicted bounding box can improve the regression accuracy, and feature refinement is a necessary process for this purpose.

- **Feature Refinement Module**

➢ Fig. 2 shows the structure of feature refinement module. Specifically, the feature map is added by two-way convolution to obtain a new feature (large kernel, LK). Only the bounding box with the highest score of each feature point is preserved in the refinement stage to increase the speed (box filtering, BF), meanwhile ensuring that each feature point corresponds to only one refined bounding box. The filtering of bounding boxes is a necessary step for feature reconstruction (FR). For each feature point of the feature map, we obtain the corresponding feature vector on the feature map according to the five coordinates of the refined bounding box (one center point and four corner points). A more accurate feature vector is obtained by bilinear interpolation. We add the five feature vectors and replace the current feature vector. After traversing the feature points, we reconstruct the whole feature map. Finally, the reconstructed feature map is added to the original feature map to complete the whole process.

- **A Derivable Approximate SkewIoU Loss**

➢ As shown in Fig. 3, each box set has the same center point, height and width. The angle difference between the two box sets is the same, but the aspect ratio is different. As a result, the smooth L1 loss value of the two sets is the same (mainly from the angle difference), but the SkewIoU is quite different. The IoU related loss is an effective regression loss function that can solve above problem and is already widely used in horizontal detection. However, the SkewIoU calculation function between two rotating boxes is underivable, which means that we cannot directly use the SkewIoU as the regression loss function. Inspired by SCRDet, we propose a derivable approximate SkewIoU loss, the multi-task loss is defined as follows:

$$L = \frac{\lambda_1}{N}\sum_{n=1}^{N} obj_n \frac{L_{reg}(v_n', v_n)}{|L_{reg}(v_n', v_n)|}|f(SkewIoU)| + \frac{\lambda_2}{N}\sum_{n=1}^{N} L_{cls}(p_n, t_n)$$

$$L_{reg}(v', v) = L_{smooth-l1}(v_\theta' - v_\theta) - IoU(v'_{\{x,y,w,h\}}, v_{\{x,y,w,h\}})$$

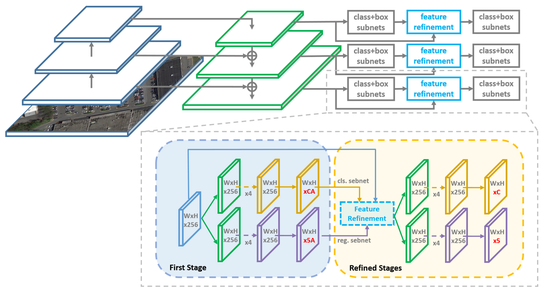

Fig1: The architecture of the proposed Refined Rotation Single-Stage Detector (RetinaNet as an embodiment).

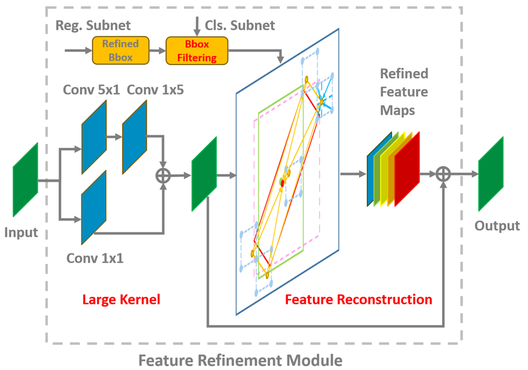

Fig2: Feature Refinement Module FRM. It mainly includes three parts: refined bounding box filtering (BF), large kernel (LK) and feature reconstruction (FR).

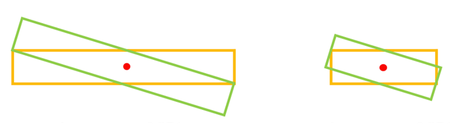

Fig3: Comparison between SkewIoU and Smooth L1.

Experiments

➢ Ablative study of each component in our method on the DOTA dataset.

Method	FRM		approximate SkewIoU loss	SV	LV	SH	mAP
	BF&FR	LK					
RetinaNet-R				64.64	71.01	68.62	62.76
RetinaNet-H				63.50	50.68	65.93	62.79
R^{3}Det*		✓		65.02	67.31	67.31	63.52
R^{3}Det	✓	✓		65.81	72.76	70.14	66.31
R^{3}Det†	✓	✓		67.45	73.98	70.27	67.66
R^{3}Det†	✓	✓	✓	68.04	72.72	76.03	69.50

➢ Ablation study for number of stages on DOTA.

#Stages	Test stage	BR	SV	LV	SH	HA	mAP
1	1	39.25	63.50	50.68	65.93	51.93	62.79
2	2	42.72	65.81	72.76	70.14	56.07	66.31
3	3	45.14	67.09	73.70	70.21	56.96	67.29
4	4	44.20	65.30	72.99	70.16	55.70	67.02
3	2 – 3	45.08	67.45	73.98	70.27	57.30	67.66

➢ Comparison between R^{3}Det* and R^{3}Det on three datasets.

Method	FRM		ICDAR205			HRSC206		UCAS-AOD
	BF&FR	LK	Recall	Precision	Hmean	mAP (07)	mAP (12)	mAP
R^{3}Det*		✓	81.64	84.97	83.27	89.14	94.98	95.03
R^{3}Det	✓	✓	83.54	86.43	84.96 (**+1.69**)	89.26 (**+0.12**)	96.01 (**+1.03**)	96.17 (**+1.14**)

➢ Experiments with our FRM with different interpolation formulas.

Align Mode	Feature Refinement Interpolation Formula	Feature Extraction	mAP (%)
FRM	$F_{lt} * A_{rb} + F_{rt} * A_{lb} + F_{rb} * A_{lt} + F_{lb} * A_{rt}$	**Bilinear**	**66.31**
FRM	$F_{lt} * A_{lt} + F_{rt} * A_{rt} + F_{rb} * A_{rb} + F_{lb} * A_{lb}$	Random Bilinear	64.28
FRM	$F_{lt} * A_{lb} + F_{rt} * A_{rb} + F_{rb} * A_{rt} + F_{lb} * A_{lt}$	Random Bilinear	64.37
FRM	$F_{lt} * 1 + F_{rt} * 0 + F_{rb} * 0 + F_{lb} * 0$	Quantification	64.02
FRM	$F_{lt} * 0 + F_{rt} * 0 + F_{rb} * 1 + F_{lb} * 0$	Quantification	64.19
deformable	–	–	63.56

➢ Detection accuracy on on DOTA.

	Method	Backbone	MS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-stage	ICN (Azimi et al. 2018)	ResNet101	✓	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
	RADet (Li et al. 2020)	ResNeXt101		79.45	76.99	48.05	65.83	65.46	74.40	68.86	89.70	78.14	74.97	49.92	64.63	66.14	71.58	62.16	69.09
	RoI-Transformer (Ding et al. 2019)	ResNet101	✓	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
	CAD-Net (Zhang, Lu, and Zhang 2019)	ResNet101		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	69.9
	Cascade-FF (Hou et al. 2020)	ResNet152		89.9	80.4	51.7	77.4	68.2	75.2	75.6	90.8	78.8	84.4	62.3	64.6	57.7	69.4	50.1	71.8
	SCRDet (Yang et al. 2019b)	ResNet101	✓	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
	FADet (Li et al. 2019)	ResNet101	✓	**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
	Gliding Vertex (Xu et al. 2020)	ResNet101		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
	Mask OBB (Wang et al. 2019)	ResNeXt101	✓	89.56	**85.95**	54.21	72.90	76.52	74.16	85.63	89.85	83.81	86.48	54.89	69.64	73.94	69.06	63.32	75.33
	FFA (Fu et al. 2020)	ResNet101	✓	90.1	82.7	54.2	75.2	71.0	79.9	83.5	90.7	83.9	84.6	61.2	68.0	70.7	76.0	63.7	75.7
	APE (Zhu, Du, and Wu 2020)	ResNeXt101		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	75.75
	CenterMap OBB (Wang et al. 2020)	ResNet101	✓	89.83	84.41	**54.60**	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03
	R^{3}Det (Lin, Feng, and Guan 2019)	ResNet101	✓	89.80	83.77	48.11	66.77	**78.76**	**83.27**	**87.84**	90.82	85.38	85.51	65.67	62.68	67.53	**78.56**	**72.62**	**76.47**
	PIoU (Chen et al. 2020)	DLA-34		80.9	69.7	24.1	60.2	38.3	64.4	64.8	90.9	77.2	70.4	46.5	37.1	57.1	61.9	64.0	60.5
	P-RSDet (Zhou et al. 2020)	ResNet101	✓	89.02	73.65	47.33	72.03	70.58	73.71	72.76	90.82	80.12	81.32	59.45	57.87	60.79	65.21	52.59	69.82
	Q-FDNet (Wei et al. 2019)	Hourglass104	✓	88.51	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 (Pan et al. 2020)	Hourglass104		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
Single-stage	R^{3}Det† (Ours)	ResNet101		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
	R^{3}Det (Ours)	ResNet152	✓	89.80	83.77	48.11	66.77	**78.76**	**83.27**	**87.84**	90.82	85.38	85.51	65.67	62.68	67.53	**78.56**	**72.62**	**76.47**

➢ Experiments with different SkewIoU functions.

Method	baseline	$-\ln(SkewIoU)$	$1 - SkewIoU$	$\exp(1 - SkewIoU) - 1$
RetinaNet-H	62.79	NAN	65.06 (**+2.27**)	65.34 (**+2.55**)
R^{3}Det†	67.66	NAN	68.97 (**+2.31**)	69.50 (**+2.84**)

➢ Evaluation on HRSC2016.

Method	Backbone	Image Size	mAP (07)	mAP (12)	Speed
R^{2}CNN (Jiang et al. 2017)	ResNet101	800*800	73.07	79.73	5fps
RC1 & RC2 (Liu et al. 2017)	VGG16	300*300	75.7	–	–
RRPN (Ma et al. 2018)	ResNet101	800*800	79.08	85.64	1.5fps
R^{2}PN (Zhang et al. 2018b)	VGG16	–	79.6	–	–
RetinaNet-H	ResNet101	800*800	82.89	89.27	14fps
RRD (Liao et al. 2018)	VGG16	384*384	84.3	–	–
RoI-Transformer (Ding et al. 2019)	ResNet101	512*800	86.20	–	6fps
Gliding Vertex (Xu et al. 2020)	ResNet101	–	88.20	–	–
DRN (Pan et al. 2020)	Hourglass104	–	–	92.70	–
SBD (Liu et al. 2019)	ResNet50	–	–	91.70	–
R^{3}Det*	ResNet101	800*800	89.14	94.98	4fps
RetinaNet-R	ResNet101	800*800	89.18	95.21	8fps
R^{3}Det	ResNet101	300*300	87.14	93.22	18fps
	ResNet101	600*600	88.97	94.61	15fps
	ResNet101	800*800	**89.26**	**96.01**	12fps
	MobileNetV2	300*300	77.16	84.31	**23fps**
	MobileNetV2	600*600	86.67	92.83	20fps
	MobileNetV2	800*800	88.71	94.45	16fps

➢ Detection accuracy on UCAS-AOD.

Method	mAP	Plane	Car
YOLOv2 (Redmon and Farhadi 2017)	87.90	96.60	79.20
R-DFPN (Yang et al. 2018b)	89.20	95.90	82.50
DRBox (Liu, Pan, and Lei 2017)	89.95	94.90	85.00
S^{2}ARN (Bao et al. 2019)	94.90	97.60	92.20
RetinaNet-H	95.47	97.34	93.60
ICN (Azimi et al. 2018)	95.67	–	–
FADet (Li et al. 2019)	95.71	**98.69**	92.72
R^{3}Det	**96.17**	98.20	**94.14**

Fig4: Visualization on DOTA datasets.