Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss Supplementary Materials

Xue Yang [1 2 3] **Junchi Yan** [1 2] **Qi Ming** [4] **Wentao Wang** [1] **Xiaopeng Zhang** [3] **Qi Tian** [3]

A. Proof of $d := \mathbf{W}(\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1); \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2))$

The entire proof process refers to this blog (Chafaï, 2010).

The Wasserstein coupling distance $\mathbf{W}$ between two probability measures μ and ν on $\mathbb{R}^n$ expressed as follows:

$$\mathbf{W}(\mu; \nu) := \inf \mathbb{E}(\|\mathbf{X} - \mathbf{Y}\|_2^2)^{1/2} \tag{1}$$

where the infimum runs over all random vectors $(\mathbf{X}, \mathbf{Y})$ of $\mathbb{R}^n \times \mathbb{R}^n$ with $\mathbf{X} \sim \mu$ and $\mathbf{Y} \sim \nu$. It turns out that we have the following formula for $d := \mathbf{W}(\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1); \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2))$:

$$d^2 = \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \mathbf{Tr}\left(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\right) \tag{2}$$

This formula interested several works (Givens et al., 1984; Olkin & Pukelsheim, 1982; Knott & Smith, 1984; Dowson & Landau, 1982). Note in particular we have:

$$\mathbf{Tr}\left((\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\right) = \mathbf{Tr}\left((\mathbf{\Sigma}_2^{1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{1/2})^{1/2}\right) \tag{3}$$

In the commutative case $\mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{\Sigma}_2\mathbf{\Sigma}_1$, Eq. 2 becomes:

$$\begin{aligned} d^2 =& \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \|\mathbf{\Sigma}_1^{1/2} - \mathbf{\Sigma}_2^{1/2}\|_F^2 \\ =&(x_1 - x_2)^2 + (y_1 - y_2)^2 + \frac{(w_1 - w_2)^2 + (h_1 - h_2)^2}{4} \\ =& l_2\text{-norm}\left(\left[x_1, y_1, \frac{w_1}{2}, \frac{h_1}{2}\right]^\top, \left[x_2, y_2, \frac{w_2}{2}, \frac{h_2}{2}\right]^\top\right) \end{aligned} \tag{4}$$

where $\|\|_F$ is the Frobenius norm. Note that both boxes are horizontal at this time, and Eq. 4 is approximately equivalent to the l_2-norm loss (note the additional denominator of 2 for w and h), which is consistent with the loss commonly used in horizontal detection. This also partly proves the correctness of using Wasserstein distance as the regression loss.

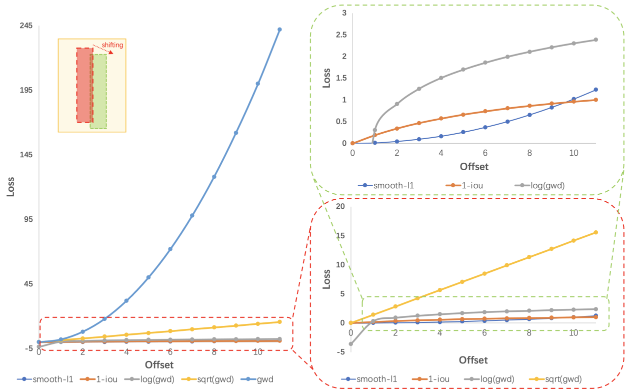

Figure 1. Different forms of GWD-based regression loss curve.

To prove Eq. 2, one can first reduce to the centered case $\mathbf{m}_1 = \mathbf{m}_2 = \mathbf{0}$. Next, if $(\mathbf{X}, \mathbf{Y})$ is a random vector (Gaussian or not) of $\mathbb{R}^n \times \mathbb{R}^n$ with covariance matrix

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{\Sigma}_2 \end{pmatrix} \tag{5}$$

then the quantity

$$\mathbb{E}(\|\mathbf{X}, \mathbf{Y}\|_2^2) = \mathbf{Tr}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2\mathbf{C}) \tag{6}$$

depends only on $\mathbf{\Gamma}$. Also, when $\mu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ and $\nu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$, one can restrict the infimum which defines W to run over Gaussian laws $\mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$ on $\mathbb{R}^n \times \mathbb{R}^n$ with covariance matrix $\mathbf{\Gamma}$ structured as above. The sole constrain on $\mathbf{C}$ is the Schur complement constraint:

$$\mathbf{\Sigma}_1 - \mathbf{C}\mathbf{\Sigma}_2^{-1}\mathbf{C}^\top \succeq 0 \tag{7}$$

The minimization of the function

$$\mathbf{C} \rightarrowtail -2\mathbf{Tr}(\mathbf{C}) \tag{8}$$

under the constraint above leads to Eq. 2. A detailed proof is given by (Givens et al., 1984). Alternatively, one may find an optimal transportation map as (Knott & Smith, 1984). It

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University [2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University [3]Huawei Inc. [4]School of Automation, Beijing Institute of Technology. Correspondence to: Junchi Yan <yangjunchi@sjtu.edu.cn>, Xue Yang <yangxue-2019-sjtu@sjtu.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Table 1. Ablation test of GWD-based regression loss form and hyperparameter on DOTA. The based detector is RetinaNet.

$1-\frac{1}{(\tau+f(d^2))}$	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=5$	$f(d^2)$	d^2
$f(\cdot)=sqrt$	68.56	**68.93**	68.37	67.77	54.27	49.11
$f(\cdot)=\log$	67.87	68.09	67.48	66.49	**69.82**	

turns out that $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$ is the image law of $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ with the linear map

$$\mathbf{x} \rightarrowtail \mathbf{m}_2 + \mathbf{A}(\mathbf{x}\mathbf{m}_1) \tag{9}$$

where

$$\mathbf{A} = \mathbf{\Sigma}_1^{-1/2}(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\mathbf{\Sigma}_1^{-1/2} = \mathbf{A}^\top \tag{10}$$

To check that this maps $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ to $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$, say in the case $\mathbf{m}_1 = \mathbf{m}_2 = \mathbf{0}$ for simplicity, one may define the random column vectors $\mathbf{X} \sim \mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $\mathbf{Y} = \mathbf{AX}$ and write

$$\begin{aligned}\mathbb{E}(\mathbf{YY}^\top) =& \mathbf{A}\mathbb{E}(\mathbf{XX}^\top)\mathbf{A}^\top \\ =& \mathbf{\Sigma}_1^{1/2}(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\mathbf{\Sigma}_1^{1/2} \\ =& \mathbf{\Sigma}_2\end{aligned} \tag{11}$$

To check that the map is optimal, one may use,

$$\begin{aligned}\mathbb{E}(\|\mathbf{X}-\mathbf{Y}\|_2^2) =& \mathbb{E}(\|\mathbf{X}\|_2^2) + \mathbb{E}(\|\mathbf{Y}\|_2^2) - 2\mathbb{E}(<\mathbf{X},\mathbf{Y}>) \\ =& \mathbf{Tr}(\mathbf{\Sigma}_1) + \mathbf{Tr}(\mathbf{\Sigma}_2) - 2\mathbb{E}(<\mathbf{X},\mathbf{AX}>) \\ =& \mathbf{Tr}(\mathbf{\Sigma}_1) + \mathbf{Tr}(\mathbf{\Sigma}_2) - 2\mathbf{Tr}(\mathbf{\Sigma}_1\mathbf{A})\end{aligned} \tag{12}$$

and observe that by the cyclic property of the trace,

$$\mathbf{Tr}(\mathbf{\Sigma}_1\mathbf{A}) = \mathbf{Tr}((\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}) \tag{13}$$

The generalizations to elliptic families of distributions and to infinite dimensional Hilbert spaces is probably easy. Some more "geometric" properties of Gaussians with respect to such distances where studied more recently by (Takatsu & Yokota, 2012) and (Takatsu & Yokota, 2012).

B. Supplementary Experiment

B.1. Improved GWD-based Regression Loss

In Tab. 1, we compare three different forms of GWD-based regression loss, including d^2, $1-\frac{1}{(\tau+f(d^2))}$ and $f(d^2)$. The performance of directly using GWD (d^2) as the regression loss is extremely poor, only 49.11%, due to its rapid growth trend (as shown on the left of Fig. 1). In other words, the regression loss d^2 is too sensitive to large errors. In contrast, $1-\frac{1}{(\tau+f(d^2))}$ achieves a significant improvement by fitting IoU loss. This loss form introduces two new hyperparameters, the non-linear function $f(\cdot)$ to transform the Wasserstein distance, and the constant τ to modulate the entire loss. From Tab. 1, the overall performance of using $sqrt$ outperforms that using $\log$, about 0.98±0.3% higher. For $f(\cdot) = sqrt$ with $\tau = 2$, the model achieves the best performance, about 68.93%. In order to further reduce the number of hyperparameters of the loss function, we directly use the GWD after nonlinear transformation ($f(d^2)$) as the regression loss. As shown in the red box in Fig. 1, $f(d^2)$ still has a nearly linear trend after transformation using the nonlinear function $sqrt$ and only achieves 54.27%. In comparison, the $\log$ function can better make the $f(d^2)$ change value close to IoU loss (see green box in Fig. 1) and achieve the highest performance, about 69.82%. In general, we do not need to strictly fit the IoU loss, and the regression loss should not be sensitive to large errors.

B.2. Training Strategies and Tricks

In order to further improve the performance of the model on DOTA, we verified many commonly used training strategies and tricks, including backbone, training schedule, data augmentation (DA), multi-scale training and testing (MS), stochastic weights averaging (SWA) (Izmailov et al., 2018; Zhang et al., 2020), multi-scale image cropping (MSC), model ensemble (ME), as shown in Tab. 2.

Backbone: Under the conditions of different detectors (RetinaNet and R^{3}Det), different training schedules (experimental groups {**#11**,**#16**}, {**#24**,**#29**}), and different tricks (experimental groups {**#26**,**#31**}, {**#28**,**#33**}), large backbone can bring stable performance improvement.

Multi-scale training and testing: Multi-scale training and testing is an effective means to improve the performance of aerial images with various object scales. In this paper, training and testing scale set to [450, 500, 640, 700, 800, 900, 1,000, 1,100, 1,200]. Experimental groups {**#3**,**#4**}, {**#5**,**#6**} and {**#11**,**#12**} show the its effectiveness, increased by 0.9%, 1.09%, and 0.58%, respectively.

Training schedule: When data augmentation and multi-scale training are added, it is necessary to appropriately lengthen the training time. From the experimental groups {**#3**,**#5**} and {**#16**,**#29**}, we can find that the performance respectively increases by 0.77% and 1.22% when the training schedule is increased from 40 or 30 epochs to 60 epochs.

Stochastic weights averaging (SWA): SWA technique has been proven to be an effective tool for improving object detection. In the light of (Zhang et al., 2020), we train our detector for an extra 12 epochs using cyclical learning rates and then average these 12 checkpoints as the final detection model. It can be seen from experimental groups {**#1**, **#2**}, {**#20**, **#21**} and {**#25**, **#26**} in Tab. 2 that we get 0.99%, 1.20% and 1.13% improvement on the challenging DOTA benchmark.

Table 2. Ablation experiment of training strategies and tricks. R-101 denotes ResNet-101 (likewise for R-18, R-50, R-152). MS, MSC, SWA, and ME represent data augmentation, multi-scale training and testing, stochastic weights averaging, multi-scale image cropping, and model ensemble, respectively. The short names for categories are defined as (abbreviation-full name): PL-Plane, BD-Baseball diamond, BR-Bridge, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle, SH-Ship, TC-Tennis court, BC-Basketball court, ST-Storage tank, SBF-Soccer-ball field, RA-Roundabout, HA-Harbor, SP-Swimming pool, and HC-Helicopter.

ID	MOETHOD	BACKBONE	SCHED.	DA	MS	MSC	SWA	ME	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	MAP$_{50}$
#1	RETINANET-GWD	R-50	20						88.49	77.88	44.07	66.08	71.92	62.56	77.94	89.75	81.43	79.64	52.30	63.52	60.25	66.51	51.63	68.93
#2							✓		88.60	78.59	44.10	67.24	70.77	62.54	79.78	88.86	81.92	80.46	57.44	64.02	62.64	66.52	55.29	69.92
#3		R-152	40	✓					89.06	83.48	49.84	65.34	74.64	67.63	82.39	88.39	84.19	84.80	63.74	61.32	66.47	70.94	67.52	73.32
#4				✓	✓				87.47	83.77	52.30	68.24	73.24	65.14	80.18	89.63	84.39	85.53	65.79	66.02	69.57	72.21	69.79	74.22
#5			60	✓					88.88	80.47	52.94	63.85	76.95	70.28	83.56	88.54	83.51	84.94	61.24	65.13	65.45	71.69	73.90	74.09
#6				✓	✓				87.12	81.64	54.79	68.74	76.17	68.39	83.93	89.06	84.51	85.99	63.33	66.68	72.60	70.63	74.17	75.18
#7				✓	✓	✓			86.14	81.59	55.33	75.57	74.20	67.34	81.75	87.48	82.80	85.46	69.47	67.20	70.97	70.91	74.07	75.35
#8				✓	✓		✓		87.63	84.32	54.83	69.99	76.17	70.12	83.13	88.96	83.19	86.06	67.72	66.17	73.47	74.57	72.80	75.94
#9				✓	✓	✓	✓		86.96	83.88	54.36	77.53	74.41	68.48	80.34	86.62	83.41	85.55	73.47	67.77	72.57	75.76	73.40	76.30
#10		–	–	✓	✓	✓	✓	✓	**89.06**	**84.32**	**55.33**	**77.53**	**76.95**	**70.28**	**83.95**	**89.75**	**84.51**	**86.06**	**73.47**	**67.77**	**72.60**	**75.76**	**74.17**	**77.43**
#11	R^{3}DET-GWD	R-101	30	✓					89.59	81.18	52.89	70.37	77.73	82.42	86.99	89.31	83.06	85.97	64.07	65.14	68.05	70.95	58.45	75.08
#12				✓	✓				89.64	81.70	52.52	72.96	76.02	82.60	87.17	89.57	81.25	86.09	62.24	65.74	68.05	74.96	64.38	75.66
#13				✓	✓		✓		89.66	82.11	52.74	71.64	75.95	83.09	86.97	89.28	85.04	86.17	65.52	63.29	72.18	74.88	63.17	76.11
#14				✓	✓	✓			89.56	81.23	53.38	79.38	75.12	82.14	86.86	88.87	81.21	86.28	65.36	65.06	72.88	73.04	62.97	76.22
#15				✓	✓	✓	✓		89.33	80.86	53.28	78.29	75.40	82.69	87.09	89.35	82.64	86.41	69.85	64.71	74.19	76.18	59.85	76.67
#16		R-152		✓					89.51	82.68	51.92	69.51	78.97	83.38	87.53	89.67	85.65	86.17	63.90	67.44	68.27	76.43	64.22	76.35
#17				✓	✓				89.55	82.28	52.39	68.30	77.86	83.40	87.48	89.56	84.27	86.14	65.38	63.25	71.33	72.36	69.21	76.18
#18				✓	✓	✓			89.62	82.27	52.35	77.30	76.95	82.53	87.20	89.08	84.58	86.21	65.21	64.46	74.99	76.30	65.19	76.95
#19		R-18	40	✓					86.63	80.12	51.98	49.67	75.73	77.54	86.10	90.05	83.22	82.31	56.05	58.86	63.30	69.06	55.07	71.05
#20				✓	✓				87.88	81.73	51.76	69.21	73.78	77.78	86.46	90.05	84.47	84.33	59.82	59.74	66.54	69.15	60.42	73.54
#21				✓	✓		✓		88.94	84.10	53.04	67.78	75.29	79.21	86.89	89.90	86.43	84.30	63.22	59.96	67.16	70.55	64.39	74.74
#22				✓	✓	✓			87.27	82.59	51.90	76.58	72.74	77.04	85.59	89.18	83.91	84.81	63.34	59.46	66.41	69.79	59.03	73.98
#23				✓	✓	✓	✓		88.38	84.75	52.63	77.35	74.29	78.53	86.32	89.12	85.73	85.13	67.84	59.48	66.88	71.59	62.58	75.37
#24		R-50	60	✓					88.82	82.94	55.63	72.75	78.52	83.10	87.46	90.21	86.36	85.44	64.70	61.41	73.46	76.94	57.38	76.34
#25				✓	✓				89.09	84.13	55.77	74.48	77.71	82.99	87.57	89.46	84.89	85.67	66.09	64.17	75.13	75.35	62.78	77.02
#26				✓	✓		✓		89.04	84.99	57.14	76.13	77.79	84.03	87.70	89.53	83.83	85.64	69.60	63.75	76.10	79.22	67.80	78.15
#27				✓	✓	✓			88.89	83.58	55.54	80.46	76.86	83.07	86.85	89.09	83.09	86.17	71.38	64.93	76.21	73.23	64.39	77.58
#28				✓	✓	✓	✓		88.43	84.33	56.91	82.19	76.69	83.23	86.78	88.90	83.93	85.73	72.07	65.67	76.76	78.37	65.31	78.35
#29		R-152		✓					88.74	82.63	54.88	70.11	78.87	84.59	87.37	89.81	84.79	86.47	66.58	64.11	75.31	78.43	70.87	77.57
#30				✓	✓				89.59	84.19	56.53	75.69	77.67	84.48	87.52	90.05	84.29	86.85	68.61	64.73	76.59	77.92	71.88	78.44
#31				✓	✓		✓		89.59	82.96	58.83	75.04	77.63	84.83	87.31	89.89	86.54	86.82	69.45	65.94	76.55	77.50	74.92	78.92
#32				✓	✓	✓			88.99	82.26	56.62	81.40	77.04	83.90	86.56	88.97	83.63	86.48	70.45	65.58	76.41	77.30	69.21	78.32
#33				✓	✓	✓	✓		89.28	83.70	59.26	79.85	76.42	83.87	86.53	89.06	85.53	86.50	73.04	67.56	76.92	77.09	71.58	79.08
#34		–	–	✓	✓	✓	✓	✓	**89.66**	**84.99**	**59.26**	**82.19**	**78.97**	**84.83**	**87.70**	**90.21**	**86.54**	**86.85**	**73.04**	**67.56**	**76.92**	**79.22**	**74.92**	**80.19**
#35	–	–	–	✓	✓	✓	✓	✓	**89.66**	**84.99**	**59.26**	**82.19**	**78.97**	**84.83**	**87.70**	**90.21**	**86.54**	**86.85**	**73.47**	**67.77**	**76.92**	**79.22**	**74.92**	**80.23**

Multi-scale image cropping: Large-scene object detection often requires image sliding window cropping before training. During testing, sliding window cropping testing is required before the results are merged. Two adjacent sub-images often have an overlapping area to ensure that the truncated object can appear in a certain sub-image completely. The cropping size needs to be moderate, too large is not conducive to the detection of small objects, and too small will cause large objects to be truncated with high probability. Multi-scale cropping is an effective detection technique that is beneficial to objects of various scales. In this paper, our multi-scale crop size and corresponding overlap size are [600, 800, 1,024, 1,300, 1,600] and [150, 200, 300, 300, 400], respectively. According to experimental groups {**#6**, **#7**} and {**#30**, **#32**}, the large object categories (e.g. GTF and SBF) that are often truncated have been significantly improved. Take group {**#6**, **#7**} as an example, GTF and SBF increased by 6.43% and 6.14%, respectively.

B.3. Comprehensive Overall Comparison

Results on DOTA: Due to the complexity of the aerial image and the large number of small, cluttered and rotated objects, DOTA is a very challenging dataset. We compare the proposed approach with other state-of-the-art methods on DOTA, as shown in Tab. 3. As far as I know, this is the most comprehensive statistical comparison of methods on the DOTA dataset. Since different methods use different image resolution, network structure, training strategies and various tricks, we cannot make absolutely fair comparisons. In terms of overall performance, our method has achieved the best performance so far, at around 80.23%.

Results on HRSC2016: The HRSC2016 contains lots of large aspect ratio ship instances with arbitrary orientation, which poses a huge challenge to the positioning accuracy of the detector. Experimental results at Tab. 4 shows that our model achieves state-of-the-art performances, about 89.85% and 97.37% in term of 2007 and 2012 evaluation metric.

References

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

Chafaï, D. Wasserstein distance between two gaussians. Website, 2010. https://djalil.chafai.net/blog/2010/04/30/wasserstein-distance-between-two-gaussians/.

Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., and Feng, J. Dual path networks. In *Advances in neural information processing systems*, pp. 4467–4475, 2017.

Chen, Z., Chen, K., Lin, W., See, J., Yu, H., Ke, Y., and Yang, C. Piou loss: Towards accurate oriented object detection in complex environments. *Proceedings of the European Conference on Computer Vision*, 2020.

Table 3. AP on different objects and mAP on DOTA. R-101 denotes ResNet-101 (likewise for R-50, R-152), RX-101 and H-104 stands for ResNeXt101 (Xie et al., 2017) and Hourglass-104 (Newell et al., 2016). Other backbone include DPN-92 (Chen et al., 2017), DLA-34 (Yu et al., 2018), DCN (Dai et al., 2017), HRNet-W48 (Wang et al., 2020a), U-Net (Ronneberger et al., 2015). MS indicates that multi-scale training or testing is used.

	Method	Backbone	MS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP$_{50}$
Two-stage methods	FR-O (Xia et al., 2018)	R-101		79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.4	52.52	46.69	44.80	46.30	52.93
	ICN (Azimi et al., 2018)	R-101	✓	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
	KARNET (Tang et al., 2020)	R-50		89.33	83.55	44.79	71.61	63.05	67.06	69.53	90.47	79.46	77.84	51.04	60.97	65.38	69.46	49.53	68.87
	RADet (Li et al., 2020b)	RX-101		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-Trans. (Ding et al., 2019)	R-101	✓	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 et al., 2019)	R-101		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
	AOOD (Zou et al., 2020)	DPN-92	✓	89.99	81.25	44.50	73.20	68.90	60.33	66.86	90.89	80.99	86.23	64.98	63.88	65.24	68.36	62.13	71.18
	Cascade-FF (Hou et al., 2020)	R-152		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., 2019)	R-101	✓	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
	SARD (Wang et al., 2019b)	R-101		89.93	84.11	54.19	72.04	68.41	61.18	66.00	90.82	87.79	86.59	65.65	64.04	66.68	68.84	68.03	72.95
	GLS-Net (Li et al., 2020a)	R-101		88.65	77.40	51.20	71.03	73.30	72.16	84.68	90.87	80.43	85.38	58.33	62.27	67.58	70.69	60.42	72.96
	FADet (Li et al., 2019)	R-101	✓	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
	MFIAR-Net (Yang et al., 2020a)	R-152	✓	89.62	84.03	52.41	70.30	70.13	67.64	77.81	90.85	85.40	86.22	63.21	64.14	68.31	70.21	62.11	73.49
	Gliding Vertex (Xu et al., 2020b)	R-101		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
	SAR (Lu et al., 2020)	R-152		89.67	79.78	54.17	68.29	71.70	77.90	84.63	90.91	**88.22**	87.07	60.49	66.95	75.13	75.28	64.29	75.28
	Mask OBB (Wang et al., 2019a)	RX-101	✓	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., 2020b)	R-101	✓	**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 et al., 2020)	RX-101		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
	F^3-Net (Ye et al., 2020)	R-152	✓	88.89	78.48	54.62	74.43	72.80	77.52	87.54	90.78	87.64	85.63	63.80	64.53	**78.06**	72.36	63.19	76.02
	CenterMap (Wang et al., 2020b)	R-101	✓	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
	CSL (Yang & Yan, 2020)	R-152	✓	**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	76.17
	MRDet (Qin et al., 2020)	R-101		89.49	84.29	55.40	66.68	76.27	82.13	87.86	90.81	86.92	85.00	52.34	65.98	76.22	76.78	67.49	76.24
	RSDet-II (Qian et al., 2021)	R-152	✓	89.93	84.45	53.77	74.35	71.52	78.31	78.12	**91.14**	87.35	86.93	65.64	65.17	75.35	79.74	63.31	76.34
	OPLD (Song et al., 2020)	R-101	✓	89.37	**85.82**	54.10	79.58	75.00	75.13	86.92	90.88	86.42	86.62	62.46	68.41	73.98	68.11	63.69	76.43
	SCRDet++ (Yang et al., 2020b)	R-101	✓	90.05	84.39	55.44	73.99	77.54	71.11	86.05	90.67	87.32	87.08	69.62	68.90	73.74	71.29	65.08	76.81
	HSP (Xu et al., 2020a)	R-101	✓	90.39	**86.23**	56.12	**80.59**	77.52	73.26	83.78	90.80	87.19	85.67	69.08	**72.02**	76.98	72.50	67.96	78.01
	FR-Est (Fu et al., 2020a)	R-101-DCN	✓	89.78	85.21	55.40	77.70	**80.26**	**83.78**	87.59	90.81	87.66	86.93	65.60	68.74	71.64	**79.99**	66.20	78.49
Single-stage methods	IENet (Lin et al., 2019)	R-101	✓	80.20	64.54	39.82	32.07	49.71	65.01	52.58	81.45	44.66	78.51	46.54	56.73	64.40	64.24	36.75	57.14
	TOSO (Feng et al., 2020)	R-101	✓	80.17	65.59	39.82	39.95	49.71	65.01	53.58	81.45	44.66	78.51	48.85	56.73	64.40	64.24	36.75	57.92
	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
	Axis Learning (Xiao et al., 2020)	R-101		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	65.98
	A^{2}S-Det (Xiao et al., 2021)	R-101		89.59	77.89	46.37	56.47	75.86	74.83	86.07	90.58	81.09	83.71	50.21	60.94	65.29	69.77	50.93	70.64
	O^2-DNet (Wei et al., 2020)	H-104	✓	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
	P-RSDet (Zhou et al., 2020)	R-101	✓	88.58	77.83	50.44	69.29	71.10	75.79	78.66	90.88	80.10	81.71	57.92	63.03	66.30	69.77	63.13	72.30
	BBAVectors (Yi et al., 2020)	R-101	✓	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	72.32
	ROPDet (Yang et al., 2020c)	R-101-DCN	✓	90.01	82.82	54.47	69.65	69.23	70.78	75.78	90.84	86.13	84.76	66.52	63.71	67.13	68.38	46.09	72.42
	HRP-Net (He et al., 2020)	HRNet-W48		89.33	81.64	48.33	75.21	71.39	74.82	77.62	90.86	81.23	81.96	62.93	62.17	66.27	66.98	62.13	72.83
	DRN (Pan et al., 2020)	H-104	✓	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
	CFC-Net (Ming et al., 2021)	R-101	✓	89.08	80.41	52.41	70.02	76.28	78.11	87.21	90.89	84.47	85.64	60.51	61.52	67.82	68.02	50.09	73.50
	R^{4}Det (Sun et al., 2020)	R-152		88.96	85.42	52.91	73.84	74.86	81.52	80.29	90.79	86.95	85.25	64.05	60.93	69.00	70.55	67.76	75.84
	R^{3}Det (Yang et al., 2021)	R-152	✓	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
	PolarDet (Zhao et al., 2020)	R-101	✓	89.65	87.07	48.14	70.97	78.53	80.34	87.45	90.76	85.63	86.87	61.64	70.32	71.92	73.09	67.15	76.64
	S^{2}A-Net-DAL (Ming et al., 2020)	R-50	✓	89.69	83.11	55.03	71.00	78.30	81.90	**88.46**	90.89	84.97	**87.46**	64.41	65.65	76.86	72.09	64.35	76.95
	R^{3}Det-DCL (?)	R-152	✓	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	77.37
	RDD (Zhong & Ao, 2020)	R-101	✓	89.15	83.92	52.51	73.06	77.81	79.00	87.08	90.62	86.72	87.15	63.96	**70.29**	76.98	75.79	72.15	77.75
	S^{2}A-Net (?)	R-101	✓	89.28	84.11	**56.95**	79.21	**80.18**	82.93	**89.21**	90.86	84.66	**87.61**	**71.66**	68.23	**78.58**	78.20	65.55	**79.15**
	GWD (Ours)	R-152	✓	89.66	84.99	**59.26**	**82.19**	78.97	**84.83**	87.70	90.21	86.54	86.85	**73.47**	67.77	76.92	**79.22**	**74.92**	**80.23**

Table 4. Detection accuracy on HRSC2016.

Method	Backbone	mAP$_{50}$ (07)	mAP$_{50}$ (12)
RC1 & RC2 (Liu et al., 2017)	VGG16	75.7	–
Axis Learning (Xiao et al., 2020)	R-101	78.15	–
TOSO (Feng et al., 2020)	R-101	79.29	–
R^{2}PN (Zhang et al., 2018)	VGG16	79.6	–
RRD (Liao et al., 2018)	VGG16	84.3	–
RoI-Trans. (Ding et al., 2019)	R-101	86.20	–
RSDet (Qian et al., 2021)	R-50	86.50	–
DRN (Pan et al., 2020)	H-104	–	92.70
CenterMap (Wang et al., 2020b)	R-50	–	92.8
SBD (Liu et al., 2019)	R-50	–	93.70
Gliding Vertex (Xu et al., 2020b)	R-101	88.20	–
OPLD (Song et al., 2020)	R-101	88.44	–
BBAVectors (Yi et al., 2020)	R-101	88.6	–
S^{2}A-Net (?)	R-101	**90.17**	95.01
R^{3}Det (Yang et al., 2021)	R-101	89.26	96.01
R^{3}Det-DCL (?)	R-101	89.46	**96.41**
FPN-CSL (Yang & Yan, 2020)	R-101	89.62	96.10
DAL (Ming et al., 2020)	R-101	89.77	–
R^{3}Det-GWD (Ours)	R-101	**89.85**	**97.37**

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773, 2017.

Ding, J., Xue, N., Long, Y., Xia, G.-S., and Lu, Q. Learning roi transformer for oriented object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2849–2858, 2019.

Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Feng, P., Lin, Y., Guan, J., He, G., Shi, H., and Chambers, J. Toso: Student'st distribution aided one-stage orientation target detection in remote sensing images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4057–4061. IEEE, 2020.

Fu, K., Chang, Z., Zhang, Y., and Sun, X. Point-based estimator for arbitrary-oriented object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020a.

Fu, K., Chang, Z., Zhang, Y., Xu, G., Zhang, K., and Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 161:294–308, 2020b.

Givens, C. R., Shortt, R. M., et al. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

He, X., Ma, S., He, L., and Ru, L. High-resolution polar network for object detection in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2020.

Hou, L., Lu, K., Xue, J., and Hao, L. Cascade detector with feature fusion for arbitrary-oriented objects in remote sensing images. In *2020 IEEE International Conference on Multimedia and Expo*, pp. 1–6. IEEE, 2020.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Knott, M. and Smith, C. S. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

Li, C., Xu, C., Cui, Z., Wang, D., Zhang, T., and Yang, J. Feature-attentioned object detection in remote sensing imagery. In *2019 IEEE International Conference on Image Processing*, pp. 3886–3890. IEEE, 2019.

Li, C., Luo, B., Hong, H., Su, X., Wang, Y., Liu, J., Wang, C., Zhang, J., and Wei, L. Object detection based on global-local saliency constraint in aerial images. *Remote Sensing*, 12(9):1435, 2020a.

Li, Y., Huang, Q., Pei, X., Jiao, L., and Shang, R. Radet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sensing*, 12(3):389, 2020b.

Liao, M., Zhu, Z., Shi, B., Xia, G.-s., and Bai, X. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5909–5918, 2018.

Lin, Y., Feng, P., and Guan, J. Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection. *arXiv preprint arXiv:1912.00969*, 2019.

Liu, Y., Zhang, S., Jin, L., Xie, L., Wu, Y., and Wang, Z. Omnidirectional scene text detection with sequential-free box discretization. *arXiv preprint arXiv:1906.02371*, 2019.

Liu, Z., Yuan, L., Weng, L., and Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, volume 2, pp. 324–331, 2017.

Lu, J., Li, T., Ma, J., Li, Z., and Jia, H. Sar: Single-stage anchor-free rotating object detection. *IEEE Access*, 8: 205902–205912, 2020.

Ming, Q., Zhou, Z., Miao, L., Zhang, H., and Li, L. Dynamic anchor learning for arbitrary-oriented object detection. *arXiv preprint arXiv:2012.04150*, 2020.

Ming, Q., Miao, L., Zhou, Z., and Dong, Y. Cfc-net: A critical feature capturing network for arbitrary-oriented object detection in remote sensing images. *arXiv preprint arXiv:2101.06849*, 2021.

Newell, A., Yang, K., and Deng, J. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pp. 483–499. Springer, 2016.

Olkin, I. and Pukelsheim, F. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

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

Qian, W., Yang, X., Peng, S., Yan, J., and Guo, Y. Learning modulated loss for rotated object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Qin, R., Liu, Q., Gao, G., Huang, D., and Wang, Y. Mrdet: A multi-head network for accurate oriented object detection in aerial images. *arXiv preprint arXiv:2012.13135*, 2020.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Song, Q., Yang, F., Yang, L., Liu, C., Hu, M., and Xia, L. Learning point-guided localization for detection in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020.

Sun, P., Zheng, Y., Zhou, Z., Xu, W., and Ren, Q. R4det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. *Image and Vision Computing*, 103:104036, 2020.

Takatsu, A. and Yokota, T. Cone structure of l2-wasserstein spaces. *Journal of Topology and Analysis*, 4(02):237–253, 2012.

Tang, T., Liu, Y., Zheng, Y., Zhu, X., and Zhao, Y. Rotating objects detection in aerial images via attention denoising and angle loss refining. *DEStech Transactions on Computer Science and Engineering*, (cisnr), 2020.

Wang, J., Ding, J., Guo, H., Cheng, W., Pan, T., and Yang, W. Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing*, 11(24):2930, 2019a.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020a.

Wang, J., Yang, W., Li, H.-C., Zhang, H., and Xia, G.-S. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020b.

Wang, Y., Zhang, Y., Zhang, Y., Zhao, L., Sun, X., and Guo, Z. Sard: Towards scale-aware rotated object detection in aerial imagery. *IEEE Access*, 7:173855–173865, 2019b.

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

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and 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.

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

Xiao, Z., Wang, K., Wan, Q., Tan, X., Xu, C., and Xia, F. A2s-det: Efficiency anchor matching in aerial image oriented object detection. *Remote Sensing*, 13(1):73, 2021.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Xu, C., Li, C., Cui, Z., Zhang, T., and Yang, J. Hierarchical semantic propagation for object detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 58(6):4353–4364, 2020a.

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.-S., and Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020b.

Yang, F., Li, W., Hu, H., Li, W., and Wang, P. Multi-scale feature integrated attention-based rotation network for object detection in vhr aerial images. *Sensors*, 20(6): 1686, 2020a.

Yang, X. and Yan, J. Arbitrary-oriented object detection with circular smooth label. In *Proceedings of the European Conference on Computer Vision*, pp. 677–694. Springer, 2020.

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., and Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8232–8241, 2019.

Yang, X., Yan, J., Yang, X., Tang, J., Liao, W., and He, T. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *arXiv preprint arXiv:2004.13316*, 2020b.

Yang, X., Yan, J., Feng, Z., and He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Yang, Z., He, K., Zou, F., Cao, W., Jia, X., Li, K., and Jiang, C. Ropdet: real-time anchor-free detector based on point set representation for rotating object. *Journal of Real-Time Image Processing*, 17(6):2127–2138, 2020c.

Ye, X., Xiong, F., Lu, J., Zhou, J., and Qian, Y. F3-net: Feature fusion and filtration network for object detection in optical remote sensing images. *Remote Sensing*, 12 (24):4027, 2020.

Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., and Metaxas, D. Oriented object detection in aerial images with box boundary-aware vectors. *arXiv preprint arXiv:2008.07043*, 2020.

Yu, F., Wang, D., Shelhamer, E., and Darrell, T. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2403–2412, 2018.

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

Zhang, H., Wang, Y., Dayoub, F., and Sünderhauf, N. Swa object detection. *arXiv preprint arXiv:2012.12645*, 2020.

Zhang, Z., Guo, W., Zhu, S., and Yu, W. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1745–1749, 2018.

Zhao, P., Qu, Z., Bu, Y., Tan, W., Ren, Y., and Pu, S. Polardet: A fast, more precise detector for rotated target in aerial images. *arXiv preprint arXiv:2010.08720*, 2020.

Zhong, B. and Ao, K. Single-stage rotation-decoupled detector for oriented object. *Remote Sensing*, 12(19):3262, 2020.

Zhou, L., Wei, H., Li, H., Zhao, W., Zhang, Y., and Zhang, Y. Arbitrary-oriented object detection in remote sensing images based on polar coordinates. *IEEE Access*, 8: 223373–223384, 2020.

Zhu, Y., Du, J., and Wu, X. Adaptive period embedding for representing oriented objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

Zou, F., Xiao, W., Ji, W., He, K., Yang, Z., Song, J., Zhou, H., and Li, K. Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. *Neural Computing and Applications*, pp. 1–14, 2020.