RankDetNet: Delving into Ranking Constraints for Object Detection

Ji Liu, Dong Li, Rongzhang Zheng, Lu Tian, Yi Shan
Xilinx Inc., Beijing, China
{jiliu1, dongl, treemann, lutian, yishan}@xilinx.com

Abstract

Modern object detection approaches cast detecting objects as optimizing two subtasks of classification and localization simultaneously. Existing methods often learn the classification task by optimizing each proposal separately and neglect the relationship among different proposals. Such detection paradigm also encounters the mismatch between classification and localization due to the inherent discrepancy of their optimization targets. In this work, we propose a ranking-based optimization algorithm for harmoniously learning to rank and localize proposals in lieu of the classification task. To this end, we comprehensively investigate three types of ranking constraints, i.e., global ranking, class-specific ranking and IoU-guided ranking losses. The global ranking loss encourages foreground samples to rank higher than background. The class-specific ranking loss ensures that positive samples rank higher than negative ones for each specific class. The IoU-guided ranking loss aims to align each pair of confidence scores with the associated pair of IoU overlap between two positive samples of a specific class. Our ranking constraints can sufficiently explore the relationships between samples from three different perspectives. They are easy-to-implement, compatible with mainstream detection frameworks and computation-free for inference. Experiments demonstrate that our RankDetNet consistently surpasses prior anchor-based and anchor-free baselines, e.g., improving RetinaNet baseline by 2.5% AP on the COCO test-dev set without bells and whistles. We also apply the proposed ranking constraints for 3D object detection and achieve improved performance, which further validates the superiority and generality of our method.

1. Introduction

Object detection is a fundamental task in computer vision with extensive subsequent research fields (e.g., instance segmentation [32] and pose estimation [40]) and wide practical applications including intelligent surveillance, autonomous driving, etc. Owing to the great advancement of convolutional neural networks (CNNs), deep learning based object detectors have achieved outstanding performance compared to the conventional hand-crafted features and classifiers. Typical methods include two-stage and one-stage anchor-based detectors. The two-stage methods [13, 36, 9, 16] first generate a set of limited candidate region proposals to distinguish foreground and background, and then classify and regress them for final detection. Two-stage methods have shown superior performance for object detection but often require large computation overhead and thus limit the applications on the resource-constrained devices. One-stage detectors [30, 34, 29] aim to identify objects from dense pre-defined anchors without the extra proposal generation step. Despite the efficiency, these one-stage detectors often suffer from the issue of foreground-background class imbalance. Recent anchor-free methods

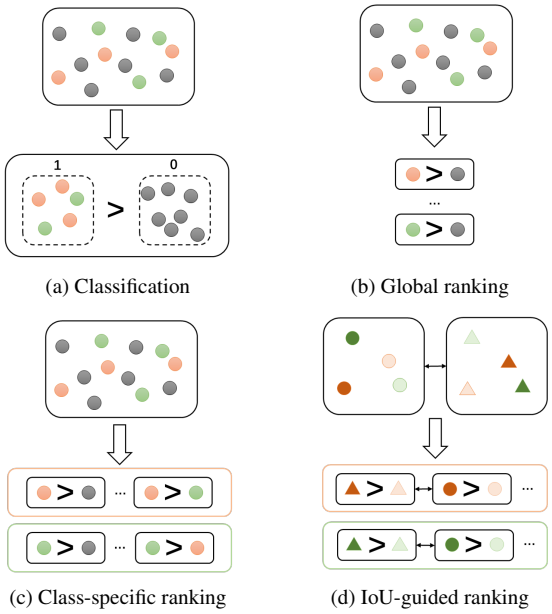


Figure 1. Illustration of the conventional classification loss and the proposed three types of pair-wise ranking losses. Gray color means background and other colors mean different object classes. Circles and Triangles indicate the object confidence and IoU overlap, respectively. Darker colors mean higher values.

[48, 24, 39, 23] attempt to detect objects by predicting points instead of enumerating possible locations, scales and aspect ratios with pre-defined anchor boxes. Most of them also follow the de facto paradigm by formulating object detection as multi-task learning of classification and localization [39, 23].

The popular focal loss [29] has been exploited to tackle the class imbalance issue by re-weighting the foreground and background samples. Although focal loss has shown improved performance for one-stage anchor-based and anchor-free detectors, it remains two limitations as follows. (1) Focal loss is designed for a single candidate sample and neglects the relationship among different samples. Treating each sample independently for optimization may lead to hard positives (i.e., positive samples with low object confidence) and hard negatives (i.e., negative samples with high object confidence), which hinders the final detection performance. (2) There exists the mismatch problem between classification and localization as the two tasks are optimized differently, which leads to accurately localized object proposals being suppressed by less accurate ones in the post-processing procedure. In fact, classification aims at distinguishing the foreground proposals from background regardless of the location information, while bounding box regression aims at localizing objects by maximizing intersection-over-union (IoU) metrics between foreground proposals and ground-truth bounding boxes. The probability learned by classification serves as *classification confidence* but localization requires *localization confidence* to learn optimal object location. Existing methods [22, 38] often adopt extra dedicated network structures to better learn localization confidence. However, they require careful designs of subnet structure or score fusion strategy and introduce additional computation cost during inference.

To alleviate these problems, we propose a ranking-based optimization algorithm to explicitly model the relationships between different pairs of proposals. In fact, classification can be viewed as a *point-wise* ranking problem where all the foreground samples are encouraged to rank higher than background, as shown in Figure 1 (a). However, we observe that it is difficult to optimize hard proposals by such point-wise ranking constraint. Therefore, we propose to replace the point-wise ranking task (i.e., classification) with optimizing two *pair-wise* ranking losses, i.e., global ranking loss and class-specific ranking loss. The global ranking loss encourages foreground samples to rank higher than background. The efficacy of our global ranking loss is similar with binary cross-entropy loss for classifying foreground and background samples but we realize it in a pair-wise manner (Figure 1 (b)). The global ranking loss treats positive samples from all the classes as foreground and does not incorporate the class information. We further propose the class-specific ranking loss to encourage that positive proposals of a specific class rank higher than negative ones (Figure 1 (c)). These negative proposals include those from other object classes as well as background. To emphasize hard foreground-background and positive-negative pairs, we design a dynamic re-weighting factor for both global and class-specific ranking losses. Moreover, to mitigate the mismatch problem between classification and localization, we propose an IoU-guided ranking loss to align each pair of object confidence scores with the corresponding pair of IoU overlap between two positive proposals of a certain class (Figure 1 (d)). It helps reduce the gap between these two subtasks and learn more localization-sensitive confidence for accurate object detection.

We evaluate our ranking-based object detection algorithm and compare with the state-of-the-art methods on the standard COCO benchmark. Our RankDetNet consistently improves the existing anchor-based and anchor-free object detectors. Especially, our RankDetNet improves RetinaNet (ResNet-50) [29] and FCOS-ATSS (ResNet-50-DCN) [46] by 2.5% and 2.4% AP on the COCO test-dev set, respectively. Without bells and whistles, our single best model obtains 48.5% AP on the COCO test-dev set under the single-scale inference setting. We also apply the proposed ranking constraints for 3D object detection and achieve performance gains of 2.02% AP over the SA-SSD baseline [15]. The results further validate the superiority and generality of our method.

The main contributions of this paper are summarized as follows. (1) We propose a ranking-based object detection algorithm by replacing the conventional classification loss with three types of pair-wise ranking losses. The proposed ranking constraints can be seamlessly integrated into existing mainstream detection framework without changing the network structure or requiring dedicated post-processing procedures. These constraints are only used for training and thus do not introduce extra computation cost for inference. (2) We view classification as a point-wise ranking problem. The global and class-specific ranking losses can serve as an effective alternative of classification in a pair-wise manner. Both constraints can reduce the number of reverse pairs during optimization. (3) We propose an IoU-guided ranking loss to address the mismatch problem between classification and localization. It helps modulate these two subtasks by aligning the object confidence score and its associated IoU overlap in a pair-wise manner, resulting in more accurate bounding box predictions. (4) Our RankDetNet achieves consistent performance improvements over 2D and 3D object detection baselines, which validates the superiority and generality of our method.

2. Related Work

Generic Object Detection. Generic object detection has achieved outstanding performance due to the advancement

of deep convolutional neural networks. Most of modern deep learning based object detectors follow the paradigm of casting object detection as classifying and regressing candidate bounding boxes in images. R-CNN [14] proposes to first generate candidate region proposals and then refine them to obtain final predictions. This two-stage detection method has been improved by a broad range of consecutive work including reducing redundant calculation of RoI features with spatial pyramid pooling [17], RoIPooling [17] or RoIAlign [16], generating region proposals by RPN [36], improving efficiency by position-sensitive score map [9] and light-head detection head [27]. While the two-stage methods have shown promising results, heavy computation loads are often entailed and applications on resource-constrained devices are limited. One-stage methods [30, 34] thus have been developed for efficient detection by directly classifying and regressing the dense pre-defined anchors without region proposal generation. In contrast to anchor mechanism, an emerging line of recent work attempts to detect objects by eliminating the requirements of pre-defined anchor boxes with different locations, scales and shapes [39, 23, 48, 49, 24]. There are different designs in these anchor-free methods for object detection such as finding object centers and regressing to their sizes [21, 48], detecting and grouping bounding box corners [24, 49], modeling all points [39] or shrunk points [23] in boxes as positive. Recently, various attempts have been made to further improve the detection performance, e.g., by addressing scale variations [28, 26], incorporating additional context information [11], refining anchor boxes [47, 41], re-weighting or mining hard samples [29, 37, 30, 5]. In this work, we propose a unified ranking-based optimization framework as a substitute for the classification task to facilitate learning accurate object detection.

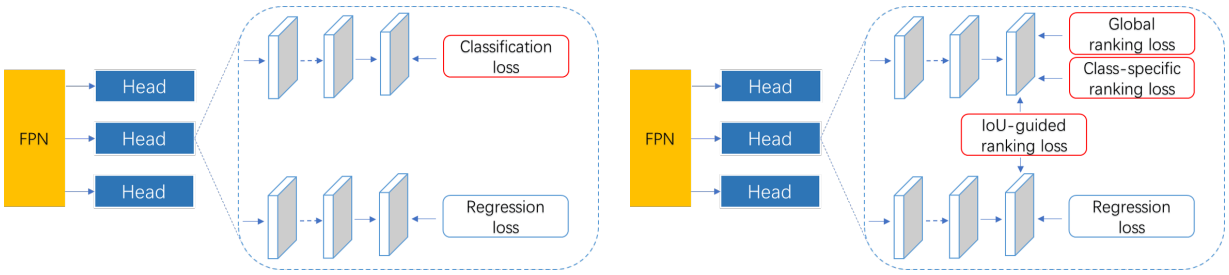
Mismatch between Classification and Localization. It is observed that the mismatch problem between classification and localization hinders the detection performance. The essence of mismatch lies in their different optimization targets. Existing methods mainly address the mismatch issue in two aspects. First, task-specific detection heads are constructed to mitigate the potential conflicts between the two subtasks. Double-Head RCNN [43] splits the branches of classification and regression right after RoIAlign and constructs them with FC layers and Conv layers respectively. Decoupling Head [31] is designed to disentangle classification and regression via the self-learned optimal feature extraction. Second, localization-sensitive scores are predicted to avoid accurately localized bounding boxes being suppressed by less accurate ones in the post-processing procedure. IoU-Net [22] introduces a branch for IoU prediction which replaces the classification score for NMS. Uncertain estimation [8, 19] is explored to learn the variance of bounding box predictions and improve the localization-sensitivity of confidence scores. Learning-to-Rank [38] utilizes an extra light-weight network to learn a ranking score of a proposal for NMS. In this work, we adopt a different way to tackle the mismatch problem by the pair-wise IoU-guided ranking loss.

Ranking Algorithms. Learning to rank has been widely used in information retrieval, search engine and recommendation system, which aims to optimize the rank of candidate pairs or sort of lists. Ranking algorithms are not only explored in classical machine learning methods (e.g., RankSVM [20]), but also developed in CNN-based methods for deep metric learning [4, 42], deep visual-semantic embedding [10], etc. Ranking algorithms can be categorized into three types, i.e, point-wise [25], pair-wise [2] and list-wise [6, 44, 3] ranking losses. Classification can be viewed as a point-wise ranking problem by shrinking samples of the same class into one point in the feature space.

In this work, we attempt to optimize object detection by replacing the classification task with the proposed pair-wise ranking constraints. Recent related methods [7, 33] also convert the classification task to a ranking problem. The main differences between prior ranking-based methods and ours are three-fold. First, the optimization targets are different. AP loss [7] targets to directly optimize the average precision metric of object detection and can be viewed as a list-wise ranking loss, while our ranking losses are optimized in a pair-wise manner. DR loss [33] optimizes the rank of the expectations of derived foreground and background distributions, while we optimize the rank of original pairs of proposals. DR loss can be viewed as a kind of global ranking loss but the algorithms on how to collect and optimize these pairs are different. Second, the implementation difficulties are different. AP loss is non-differentiable and non-convex and hence needs a tailored approximate algorithm for optimization. DR loss also needs complicated regularization and smooth approximation algorithms to obtain the constrained distribution. Differently, our ranking constraints are easier to implement and can be directly trained by the standard gradient descent methods without extra complicated optimization strategies. Third, [7, 33] only optimize the ranking problem for classification while we also constrain the relationship between classification and localization with the explicit IoU-guided ranking loss.

3. Proposed Approach

In the object detection baseline method (Figure 2 (a)), two subtasks of classification and localization are optimized in a multi-task manner. Given a set of candidate proposals from an image, the classification task aims to identify the foreground proposals from background ones. The classification task can be optimized by:



(a) Baseline detection network

(b) Our RankDetNet

Figure 2. (a) The baseline detection network consists of two subtasks of classification and localization. (b) The proposed RankDetNet replaces the classification loss with the three types of pair-wise ranking losses.

$$\min_{\theta} \sum_{i}^{N} \Big(\sum_{p_i \in \mathcal{P}_i} \mathcal{L}_{\text{cls}}(f_\theta(p_i)) + \sum_{n_i \in \mathcal{N}_i} \mathcal{L}_{\text{cls}}(f_\theta(n_i)) \Big) \quad (1)$$

where θ denotes the network parameter to be learned, N is the number of total training images, $\mathcal{L}_{\text{cls}}$ is the loss function for classification, $f_\theta(\cdot)$ predicts the object confidence for each candidate. p_i and n_i indicate the positive and negative samples from the candidate sets of $\mathcal{P}_i$ and $\mathcal{N}_i$ in the i-th image, respectively. The conventional cross-entropy loss encounters the class imbalance problem as $|\mathcal{N}| \gg |\mathcal{P}|$ during optimization. The focal loss alleviates such issue with a re-weigting scheme but does not consider the relationship among samples as it computes the loss for each sample separately. In this work, we propose a different way for object detection with a ranking-based learning framework. For brevity, we will omit the index of image in the subsequent sections.

3.1. Global Ranking Loss

In order to model the relationships between foreground and background samples, we propose a global ranking loss to encourage foreground samples to rank higher than background ones. Our global ranking loss can be formulated as:

$$\min_{\theta} \sum_{n \in \mathcal{N}} \sum_{p \in \mathcal{P}} z \cdot \mathcal{L}_{\text{rank}}(f_\theta(n) - f_\theta(p)) \quad (2)$$

where z is a dynamic re-weighting factor:

$$z(n,p) = \frac{\mathcal{L}_{\text{rank}}(f_\theta(n) - f_\theta(p))}{\sum_{n' \in \mathcal{N}} \sum_{p' \in \mathcal{P}} \mathcal{L}_{\text{rank}}(f_\theta(n') - f_\theta(p'))} \quad (3)$$

$\mathcal{L}_{\text{rank}}$ is a non-negative monotonically-increasing function and we use the exponential form $\exp(\cdot)$ in our method. To construct the foreground-background pairs, we collect all the positive proposals from all the object classes as foreground. We design two schemes to construct negative proposals: (1) We apply OHEM [30] to mine hard negatives as background. The ratio between foreground and background samples is enforced to 1:3. (2) We sort all the background samples based on the object confidence scores and divide them into multiple groups at equal intervals of scores. We then use the average scores of each group to compute the loss. The dynamic re-weighting factor z is designed to emphasize the learning of foreground-background reverse pairs (i.e., object confidence of a background sample is higher than a foreground one) with the large score gap. We constrain no gradient back-propagating along the variable during training.

Compared to binary classification which can be viewed as a point-wise ranking problem, our global ranking loss is optimized in a pair-wise manner differently. Since each pair consists of a foreground sample and a background one, it is well balanced with respect to the foreground and background classes during optimization. We observe that the foreground-background reverse pairs (i.e. hard pairs) can be reduced by training the proposed global ranking loss.

3.2. Class-Specific Ranking Loss

The global ranking loss aims to rank all the foreground samples from all the object classes higher than background. It does not incorporate the class information to distinguish different object classes. Thus, we propose a class-specific ranking loss to encourage positive samples from a specific class to rank higher than negative ones. In this case, the negative samples can be collected from the other object classes as well as background. Our class-specific ranking loss can be formulated as:

$$\min_{\theta} \sum_{c=0}^{C} \sum_{n \in \mathcal{N}_c} \sum_{p \in \mathcal{P}_c} z_c \cdot \mathcal{L}_{\text{rank}}(f_\theta(n) - f_\theta(p)) \quad (4)$$

where z_c is a dynamic re-weighting factor:

$$z_c(n,p) = \frac{\mathcal{L}_{\text{rank}}(f_\theta(n) - f_\theta(p))}{\sum_{n' \in \mathcal{N}_c} \sum_{p' \in \mathcal{P}_c} \mathcal{L}_{\text{rank}}(f_\theta(n') - f_\theta(p'))} \quad (5)$$

Here, C is the total number of object classes, $\mathcal{N}_c$ and $\mathcal{P}_c$ indicate the sets of candidate negative and positive proposals for the c-th class, z_c is a dynamic re-weighting factor

for the c-th class. For each class, we collect all the positive samples to compute the loss. To mine negative samples, we apply "OHEM 1:3" or "group" strategies which are similar with our global ranking loss. Training with the class-specific ranking loss can reduce the positive-negative reverse pairs and mitigate the positive-negative sample imbalance problem for each class. Our global and class-specific ranking losses are complementary to each other and serve as an alternative of the classification problem in a pair-wise manner.

3.3. IoU-Guided Ranking Loss

To tackle the mismatch between classification confidence and localization confidence, we propose an IoU-guided ranking loss to encourage one positive sample with a larger IoU overlap to rank higher than another one with smaller IoU. To this end, we align each pair of confidence scores with the associated pair of IoU overlap between two positive samples of a specific class. The optimization objective can be formulated in a pair-wise manner:

$$\begin{aligned} &\min_\theta \sum_{c=0}^{C} \sum_{k=0}^{K} \sum_{p_i \in \mathcal{P}_c^k} \sum_{p_j \in \mathcal{P}_c^k} \mathcal{L}(\theta), \\ &\mathcal{L}(\theta) = \mathcal{L}_{\text{rank}}(-\alpha \cdot (f_\theta(p_i) - f_\theta(p_j))(s_\theta(p_i) - s_\theta(p_j))) \end{aligned} \tag{6}$$

where K is the number of levels of detection head (e.g., 5 levels for RetinaNet with ResNet-50-FPN) and $\alpha > 0$ is a hyper-parameter to control the loss value. $\mathcal{P}_c^k$ indicates the set of candidate positive proposals from the c-th class and the k-th pyramid level. $s_\theta(\cdot)$ denotes the IoU overlap between the candidate proposal and the nearby ground-truth bounding box. In contrast to the global and class-specific ranking losses where samples can be collected across different pyramid levels, we collect samples from each separate level to compute the IoU-guided ranking loss. This is because distributions of IoU overlap are different in different pyramid levels. Specifically, the deeper level contains larger proposals which tend to generate higher overlap value with GT according to the IoU metric and vice versa.

During the SGD optimization process, we constrain no gradient back-propagating along the IoU score term with the lower object confidence score. The other IoU score term as well as two confidence score terms are normally optimized. Without loss of generality, suppose $f_\theta(p_i) > f_\theta(p_j)$, if $s_\theta(p_i) < s_\theta(p_j)$, we freeze $s_\theta(p_j)$ and optimize $s_\theta(p_i)$ to he higher than $s_\theta(p_j)$; if $s_\theta(p_i) > s_\theta(p_j)$, we still freeze $s_\theta(p_j)$ and continue to optimize $s_\theta(p_i)$. If $s_\theta(p_j)$ is not frozen, the loss still can drop by decreasing $s_\theta(p_j)$, i.e., decreasing the IoU overlap between positive samples and GT, which causes an unexpected optimization direction. Our IoU-guided ranking loss is different with IoU loss [45]. The IoU loss handles each sample separately and only optimizes the IoU score term, while we handle samples in a pair-wise manner and optimize object confidence and location simultaneously.

The proposed IoU-guided ranking loss can help learn localization-sensitive object confidence and help reduce the misalignment between proposal classification and location regression, leading to more accurate bounding box predictions.

3.4. RankDetNet Detector

The proposed ranking constraints can be easily integrated into the existing baseline detection network by replacing the classification task with the proposed ranking losses, as shown in Figure 2 (b). We simply combine the three types of ranking losses with the regression loss by weights of 1:1:1:1 for training and follow the baseline detectors for testing. Our RankDetNet does not need to tweak the original network and do not introduce extra computation cost for inference.

4. Experiments

4.1. Experimental Settings

Datasets. All the experiments are conducted on the MS COCO 2017 dataset with 80 object classes for object detection. We train the detector on the training set containing 118K images and evaluate the performance on the validation set with 5K images. We also report results on the COCO test-dev set. The standard COCO AP is used as the evaluation metrics.

Training Details. We use pre-trained networks on ImageNet as backbone (e.g., ResNet-50 [18]) and construct a feature pyramid network [28] on top for detection. Images are resized to a maximum scale of 800×1333 without changing the aspect ratio. The network is optimized with Stochastic Gradient Descent (SGD) with momentum of 0.9, weight decay of 0.0001 and batch size of 16. We first use focal loss to train the network for providing initialized weights and then train our ranking losses by 12 epochs. Such warm-up scheme can help improve the stability of training. The initial learning rate is set as 0.01 and decayed by 10× at 8-th and 11-th epochs. For the OHEM [30] strategy, we use all the positives and collect 3× hard negatives. For the group strategy, we set numbers of groups as 15 in the global ranking loss and 3 for each class in the class-specific ranking loss. In the IoU-guided ranking loss, we set $\alpha = 1.0$ for all the experiments. Our experiments are conducted on Pytorch and MMdetection-1.0 platforms with 4 V100 GPUs. Our RankDetNet cost 8.95G memory per GPU and 4h per epoch with the RetinaNet (ResNet-50) baseline for training.

Inference Details. For inference, we pre-process the input image using the same procedure as the training phase. The network outputs the predicted bounding boxes and their associated class probabilities. Following [29], we first filter out a large number of background bounding boxes by setting an IoU threshold as 0.05 and then generate top 1,000

Methods	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet	ResNet-50	35.4	55.3	37.7	20.1	39.1	46.0
RetinaNet + RankDetNet	ResNet-50	37.8	57.1	40.7	20.8	41.0	50.1
RetinaNet	ResNet-101	37.5	57.6	40.3	21.2	42.0	49.2
RetinaNet + RankDetNet	ResNet-101	39.4	59.1	42.3	21.6	43.2	52.3
RetinaNet	ResNeXt-64×4d-101	39.7	60.5	42.8	23.4	44.1	51.8
RetinaNet + RankDetNet	ResNeXt-64×4d-101	41.6	61.4	44.9	23.0	45.6	55.1
FCOS-ATSS	ResNet-50	39.2	57.2	42.5	23.2	43.2	50.9
FCOS-ATSS + RankDetNet	ResNet-50	40.7	58.4	44.0	23.8	44.8	52.8
FCOS-ATSS	ResNet-50-DCN	43.9	62.1	47.6	26.1	48.0	57.9
FCOS-ATSS + RankDetNet	ResNet-50-DCN	45.2	63.4	49.1	26.7	49.1	60.5
FCOS-ATSS	ResNeXt-64×4d-101-DCN	47.1	66.1	51.0	29.5	51.2	61.7
FCOS-ATSS + RankDetNet	ResNeXt-64×4d-101-DCN	**48.1**	**66.7**	**52.5**	**29.8**	**52.3**	**63.4**

Table 1. Detection performance comparisons (%) on the COCO 2017 validation set. For RetinaNet and FCOS-ATSS baselines, we report the performance with our re-implementation. For fair comparisons, all the results are obtained in the same single-scale training and inference settings.

Methods	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Anchor-Based Detectors:							
SSD513 [30]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
YOLOv3 (608×608) [35]	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9
Faster R-CNN w/ FPN [28]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
PISA [5]	ResNet-50	37.3	56.5	40.3	20.3	40.4	47.2
AP-Loss [7]	ResNet-101	37.4	58.6	40.5	17.3	40.8	51.9
DR-Loss [33]	ResNet-50	37.6	-	-	-	-	-
Mask R-CNN [16]	ResNet-101	38.2	60.3	41.7	20.1	41.1	50.2
RetinaNet [29]	ResNet-50	35.7	55.0	38.5	18.9	38.9	46.3
RetinaNet [29]	ResNet-101	37.8	57.5	40.8	20.2	41.1	49.2
Anchor-Free Detectors:							
FoveaBox [23]	ResNet-50	37.1	57.2	39.5	21.6	41.4	49.1
FCOS [39]	ResNet-50	37.1	55.9	39.8	21.3	41.0	47.8
FSAF [50]	ResNet-50	37.2	57.2	39.4	21.0	41.2	49.7
ExtremeNet [49]	Hourglass-104	40.2	55.5	43.2	20.4	43.2	53.1
CornerNet [24]	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
CenterNet-HG [48]	Hourglass-104	42.1	61.1	45.9	24.1	45.5	52.8
FCOS-ATSS [46]	ResNet-50	39.3	-	-	-	-	-
FCOS-ATSS [46]	ResNet-50-DCN	43.0	-	-	-	-	-
Ours:							
RetinaNet + RankDetNet	ResNet-50	38.2	57.8	41.2	20.9	40.8	48.4
RetinaNet + RankDetNet	ResNet-101	40.0	59.7	43.2	21.9	43.0	50.6
RetinaNet + RankDetNet	ResNeXt-64×4d-101	42.1	62.1	45.4	23.8	45.2	53.2
FCOS-ATSS + RankDetNet	ResNet-50	41.1	58.9	44.9	24.3	43.9	51.2
FCOS-ATSS + RankDetNet	ResNet-50-DCN	45.4	63.6	49.5	26.7	48.5	58.1
FCOS-ATSS + RankDetNet	ResNeXt-64×4d-101-DCN	48.5	67.1	52.8	29.4	51.8	61.6

Table 2. Detection performance comparisons (%) on the COCO 2017 test-dev set. For fair comparisons, all the results are obtained in the same single-scale inference settings.

detection candidates per feature pyramid level. Then, we apply soft Non-Maximum Suppression (soft-NMS) [1] with IoU threshold of 0.6 to yield the final detection results.

4.2. RankDetNet for 2D Object Detection

Comparisons to the Anchor-Based and Anchor-Free Baselines. Table 1 compares our RankDetNet with the popular anchor-based and anchor-free object detectors on the validation set. First, taking RetinaNet [29] as the anchor-based detection baseline, we achieve consistent performance improvements with different backbones, e.g., surpassing RetinaNet (ResNet-50) by 2.4% AP. Second, taking FCOS-ATSS [46] as the anchor-free detection baseline, our RankDetNet also brings further improvements, leading to 48.1% AP using a single ResNeXt-64×4d-101-DCN model (without multi-scale training/testing and without model ensemble). Third, for both anchor-based and anchor-free baselines, our method achieves larger gains for larger objects and remarkably enhance the performance according to a more strict IoU overlap criterion. For example, using the same ResNet-50 backbone, we improve AP_L by +4.1% and AP_{75} by +3.0% over the RetinaNet baseline. Similarly,

we improve AP_L by +1.9% and AP_{75} by +1.5% over the FCOS-ATSS (ResNet-50) method.

Comparisons to the State-of-the-Arts. Table 2 compares our RankDetNet with the state-of-the-art object detectors on the test-dev set. Specifically, our method outperforms the one-stage RetinaNet baseline with ResNet-50 by 2.5% AP. We also perform favorably against prior two-stage methods, e.g., surpassing Faster R-CNN and Mask R-CNN by 3.8% and 1.8% AP, respectively, with the same ResNet-101 backbone. Our RankDetNet also achieves competitive performance compared to the anchor-free methods, e.g., improving FCOS-ATSS with ResNet-50-DCN by 2.4% AP. Equipped with better network backbones, we can further enhance the detection performance. With the same multi-scale training strategy and the same ResNeXt-64×4d-101-DCN backbone, our method outperforms FCOS-ATSS by 2.3% (47.7% vs. 50.0%). Compared to the other ranking based detection methods, our RankDetNet outperforms AP loss [7] by 2.6% AP with ResNet-101 and performs slightly better than DR loss [33] with ResNet-50. It is worth noting that our ranking losses are easier to implement and do not require extra complicated optimization strategies.

Analysis of Reserve Pairs. To better understand the effect of our ranking constraints, we define three types of reverse pairs and compare the ratios of reverse pairs to all pairs after training in Table 3. We observe that training with the global ranking loss or class-specific ranking loss only is not sufficient, which may incur higher reverse pairs than focal loss. Combining these two losses can notably reduce the foreground-background / positive-negative reverse pairs. By imposing the additional IoU-guided ranking loss, the amount of reverse pairs can be further reduced. It encourages that positive samples with larger IoU overlap with GT have higher object confidence scores. These results demonstrate the efficacy of our ranking constraints for better optimization.

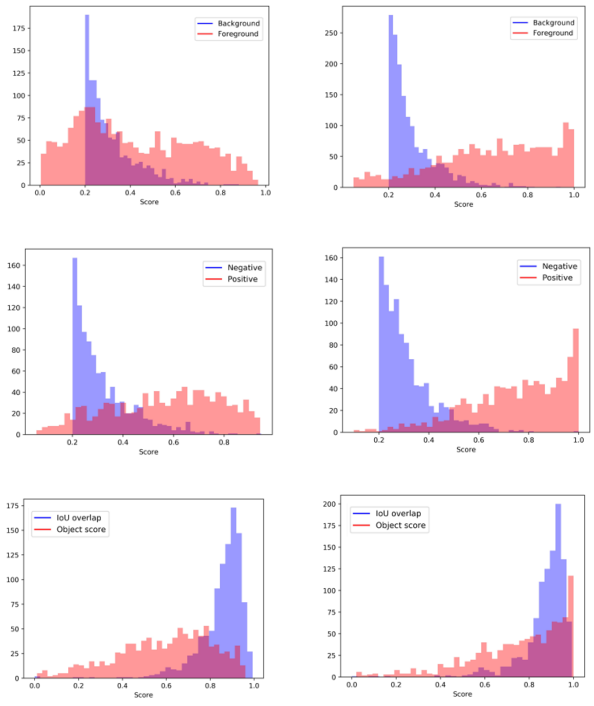
Analysis of Sample Distribution. Figure 3 shows sample distributions for the focal loss baseline and our method. For the foreground or positive samples, higher object confidence scores can be learned by our method. The reduced distribution overlap between foreground and background by our method implies that our ranking losses can better distinguish samples to help detection optimization. Similar conclusions can be made for the overlap between positive and negative distribution. Compared with IoU overlap and object confidence scores for positive samples, Figure 3 shows our method obtains more consistent distributions and stronger correlation.

4.3. Ablation Study

We conduct ablation studies to examine each contribution of algorithmic components with the RetinaNet (ResNet-50) baseline in Table 4. Combining the global and class-specific ranking losses can largely improve the performance over the method of using either one only. We hypothesize that optimization with only the individual global or class-specific ranking loss is insufficient for distinguishing objects and background. The proposed dynamic re-weighting factors aim to emphasize hard pairs during training, which are beneficial for performance boost in the global and class-specific losses (36.6% vs. 37.3%). For the IoU-guided ranking loss, we find that constructing pairs from

Loss	F-B	P-N	P-P
Focal	0.33	0.28	0.44
Global	0.33	0.29	0.43
Class-specific	0.39	**0.21**	0.44
Global + class-specific	0.27	0.23	0.43
Global + class-specific + IoU-guided	**0.26**	0.22	**0.41**

Table 3. Ratios of reverse pairs for different losses. F-B: foreground-background reverse pair where foreground has lower score. P-N: positive-negative reverse pair where the positive sample of a specific class has lower score. P-P: positive-positive reverse pair where the one positive sample with larger IoU overlap with GT has lower object confidence than the other.



(a) Focal loss (b) Our method

Figure 3. Comparisons of sample distribution. We plot foreground and background distribution (first row), positive and negative distribution for the person class (second row) and distribution of IoU overlap and object confidence score for the person class (third row). We select background / negative samples with score > 0.2 for better visualization.

Exp.	Global	Class-specific	IoU-guided	Dynamic re-weighting	Pyramid level	IoU BP	Negative mining	AP
(1)	✓	✗	✗	✓	✗	✗	OHEM	34.2
(2)	✗	✓	✗	✓	✗	✗	OHEM	32.7
(3)	✓	✓	✗	✓	✗	✗	OHEM	36.6
(4)	✗	✗	✓	✗	Separate	$s_\theta(p_i)$	OHEM	36.2
(5)	✓	✓	✓	✓	Merge	✗	OHEM	36.7
(6)	✓	✓	✓	✓	Separate	✗	OHEM	37.0
(7)	✓	✓	✓	✗	Separate	$s_\theta(p_i)$	OHEM	36.6
(8)	✓	✓	✓	✓	Separate	$s_\theta(p_i)$ & $s_\theta(p_j)$	OHEM	35.4
(9)	✓	✓	✓	✓	Separate	$s_\theta(p_i)$	OHEM	37.3
(10)	✓	✓	✓	✓	Separate	$s_\theta(p_i)$	Group	**37.8**

Table 4. Ablation studies of the proposed RankDetNet with the RetinaNet (ResNet-50) baseline on the COCO 2017 validation set. "Merge" and "Separate" mean computing IoU-guided ranking loss from merged or separate pyramid levels, respectively. For IoU BP, we suppose $f_\theta(p_i) > f_\theta(p_j)$, then we freeze $s_\theta(p_j)$ and only optimize $s_\theta(p_i)$. Exp. (4) is trained with focal loss + IoU-guided ranking loss.

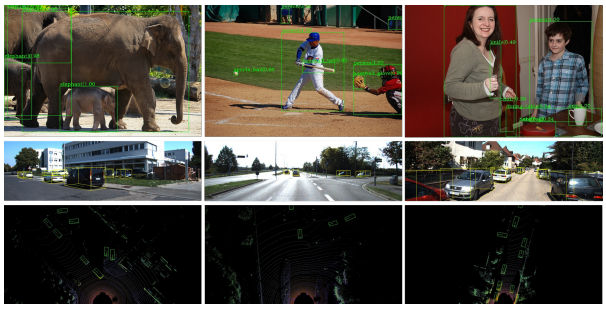


Figure 4. Sample 2D and 3D detection results by our RankDetNet.

Loss	AP
Exponential	**37.3**
Sigmoid	37.2
Logistic	**37.3**

(a)

α	AP
0.2	37.0
0.5	37.0
1.0	**37.3**
5.0	36.0

(b)

Table 5. Ablation studies of (a) different ranking loss functions in our RankDetNet and (b) hyper-parameter α in the IoU-guided ranking loss on the COCO 2017 validation set. We use RetinaNet (ResNet-50) as baseline and adopt OHEM for negative mining in these experiments.

each separate pyramid level works better than merging them across different levels. The results validate our observations that distributions of IoU overlap scores are different in different pyramid levels. Table 4 also shows that constraining no gradient back-propagating along the IoU score term with the lower object confidence score is crucial for the IoU-guided ranking loss. This is because decreasing the IoU overlap between positive samples and GT will lead to an unexpected optimization direction. For negative mining, the group strategy can bring further improvements compared to OHEM (37.3% vs. 37.8%). We also test different ranking functions $\mathcal{L}_{\text{rank}}(\cdot)$ and obtain similar results, as shown in Table 5 (a). Table 5 (b) presents the effect of α in the IoU-guided ranking loss. We find that $\alpha = 1.0$ works best and choose this value for all the experiments.

4.4. RankDetNet for 3D Object Detection

Our RankDetNet framework can be readily extended to the 3D object detection task. We use the standard KITTI [12] benchmark which contains 7,481 annotated scans of point cloud (3,712 for training and 3,769 for validation) in our experiments. We adopt one of the state-of-the-art 3D detection methods of SA-SSD [15] as our baseline. SA-SSD designs a backbone network with 3D Conv layers to extract multi-scale features from point cloud and also employs a focal loss in the detection head to predict 3D bounding box. By substituting the focal loss for proposal classification with our ranking losses, we achieve 86.32% AP@(IoU=0.7) for car on the *moderate* set with a gain of 2.02% compared to the SA-SSD baseline. The results validate the effectiveness of the proposed ranking constraints for 3D object detection and further exhibit the generality of our method.

5. Conclusion

In this work, we proposed a unified ranking-based optimization algorithm for object detection with three types of ranking constraints, i.e., global ranking, class-specific ranking and IoU-guided ranking losses. Our ranking constraints can sufficiently explore the relationships between samples from three different perspectives. They are easy-to-implement and can be seamlessly integrated into mainstream detection frameworks, without introducing extra computation cost for inference. Experiments show that our RankDetNet consistently improves the state-of-the-art anchor-based and anchor-free 2D detection baselines and 3D detection methods, validating the superiority and generality of our method. We expect that the proposed ranking constraints will inspire new insights for object detection and other similar tasks.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, 2017. 6

[2] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *ICML*, 2005. 3

[3] Christopher J Burges, Robert Ragno, and Quoc V Le. Learning to rank with nonsmooth cost functions. In *NeurIPS*, 2007. 3

[4] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *CVPR*, 2019. 3

[5] Yuhang Cao, Kai Chen, Chen Change Loy, and Dahua Lin. Prime sample attention in object detection. In *CVPR*, 2020. 3, 6

[6] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, 2007. 3

[7] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with ap-loss. In *CVPR*, 2019. 3, 6, 7

[8] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *ICCV*, 2019. 3

[9] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NeurIPS*, 2016. 1, 3

[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 3

[11] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 3

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 8

[13] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 1

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 3

[15] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In *CVPR*, 2020. 2, 8

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 3, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI*, 37(9):1904–1916, 2015. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[19] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *CVPR*, 2019. 3

[20] Ralf Herbrich. Large margin rank boundaries for ordinal regression. *Advances in large margin classifiers*, pages 115–132, 2000. 3

[21] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 3

[22] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, 2018. 2, 3

[23] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, and Jianbo Shi. Foveabox: Beyond anchor-based object detector. *arXiv preprint arXiv:1904.03797*, 2019. 2, 3, 6

[24] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 2, 3, 6

[25] Ping Li, Qiang Wu, and Christopher J Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. In *NeurIPS*, 2008. 3

[26] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *ICCV*, 2019. 3

[27] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*, 2017. 3

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3, 5, 6

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1, 2, 3, 5, 6

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 1, 3, 4, 5, 6

[31] Yu Liu, Guanglu Song, Yuhang Zang, Yan Gao, Enze Xie, Junjie Yan, Chen Change Loy, and Xiaogang Wang. 1st place solutions for openimage2019–object detection and instance segmentation. *arXiv preprint arXiv:2003.07557*, 2020. 3

[32] Pedro OO Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NeurIPS*, 2015. 1

[33] Qi Qian, Lei Chen, Hao Li, and Rong Jin. Dr loss: Improving object detection by distributional ranking. In *CVPR*, 2020. 3, 6, 7

[34] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 1, 3

[35] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 6

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2015. 1, 3

[37] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 3

[38] Zhiyu Tan, Xuecheng Nie, Qi Qian, Nan Li, and Hao Li. Learning to rank proposals for object detection. In *ICCV*, 2019. 2, 3

[39] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 2, 3, 6

[40] A Toshev, Christian Szegedy, and G DeepPose. Human pose estimation via deep neural networks. In *CVPR*, 2014. 1

[41] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *CVPR*, 2019. 3

[42] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In *CVPR*, 2019. 3

[43] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Double-head rcnn: Rethinking classification and localization for object detection. In *CVPR*, 2020. 3

[44] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *ICML*, 2008. 3

[45] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: Anadvancedobjectdetectionnetwork. In *ACM MM*, 2016. 5

[46] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, 2020. 2, 6

[47] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018. 3

[48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2, 3, 6

[49] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019. 3, 6

[50] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *CVPR*, 2019. 6