BALANCED RANKING AND SORTING FOR CLASS INCREMENTAL OBJECT DETECTION

Bo Cui[1,2⋆] *Hui Qu*[5] *Xuhui Huang*[5] *Shan Yu*[1,3,4]

[1] Brainnetome Center and NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] School of Future Technology, University of Chinese Academy of Sciences
[4] CAS Center for Excellence in Brain Science and Intelligence Technology
[5] X-Lab, the Second Academy of CASIC, Beijing, China

ABSTRACT

Class incremental learning has drawn much attention recently. Although many algorithms have been proposed for class incremental image classification, developing object detectors which can learn incrementally is still a challenge. Existing methods rely on knowledge distillation to achieve class incremental object detection (CIOD), which suffer from performance tradeoff between old and new classes. In this paper, we propose balanced ranking and sorting (BRS), to tackle the catastrophic forgetting and data imbalance problems for CIOD. Specifically, ranking & sorting with pseudo ground truths (RSP) and ranking & sorting transfer (RST) are developed to preserve the learned knowledge from the old model while learning new classes, in an unified framework. To mitigate the data imbalance problem, gradient rebalancing is performed with specific sample pairs. We demonstrate the effectiveness of our approach with extensive experiments on PASCAL VOC and COCO datasets, in which significant improvement over state-of-the-art methods is achieved.

***Index Terms*—** Class incremental learning, object detection, rank and sort, data imbalance

1. INTRODUCTION

Object detection is one of the most important computer vision tasks and has broad applications. Modern object detectors [1–5] developed using deep convolutional networks can localize and classify object regions simultaneously. These methods have achieved state-of-the-art results on datasets such as PASCAL VOC [6] and COCO [7]. However, most of these models can only detect classes which have been annotated for training. As humans learn in a life-long manner, it's appealing to enable detectors to learn continually, in other words, sequentially learn to detect new classes without access to the past annotated data. Most of existing studies on class incremental learning focus on image classification task. There are relatively few methods designed for incremental object detection.

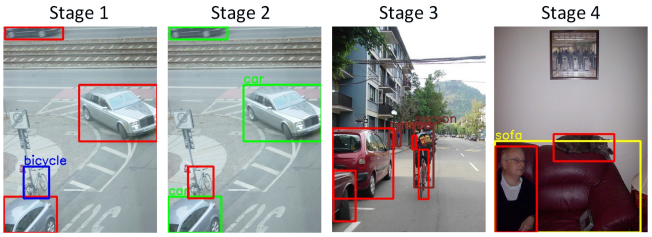

Fig. 1. A 4-stage learning setting for CIOD. In stage $t > 1$, a few training images may have been used in earlier stages. However, only the newest classes (such as car in stage 2) are annotated. Red boxes are the instances without annotations

Recently, an incremental learning framework for object detection, ILOD, is proposed [8]. The detection problem is converted to a classification task on pre-computed region proposals with Fast R-CNN [1]. ILOD uses distillation-based method [9] to preserve the performance on old classes. RKT [10] improves ILOD via proposal selection and relation guided knowledge transfer. However, the region proposals are obtained using EdgeBoxes [11] or MCG [12], with inferior quality than the ones generated with modern region proposal network (RPN) [2]. More recent related works are [13] and [14] which apply knowledge distillation (KD) to Faster-RCNN [2] and anchor-free detectors. It's challenging to learn object detectors incrementally without catastrophic forgetting.

In this study, we propose a balanced ranking and sorting method to tackle the class incremental object detection problem. First, the CIOD task setting is introduced. As shown in Fig. 1, there are several sequential learning stages that new classes are added in each stage. The training set for each learning stage contains all the images that have at least one instance of the new classes, and only the new classes are annotated. The distribution of training data in the final stage of a 4-stage learning setting (here refers to learning 20 classes of COCO dataset every stage) is shown in Fig. 2, in which a few classes contribute to most of the training samples, while most of the old classes are under-represented. So the key to solve

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB32040200) and Beijing Academy of Artificial Intelligence. ⋆The corresponding author is Bo Cui, E-mail: bo.cui@nlpr.ia.ac.cn

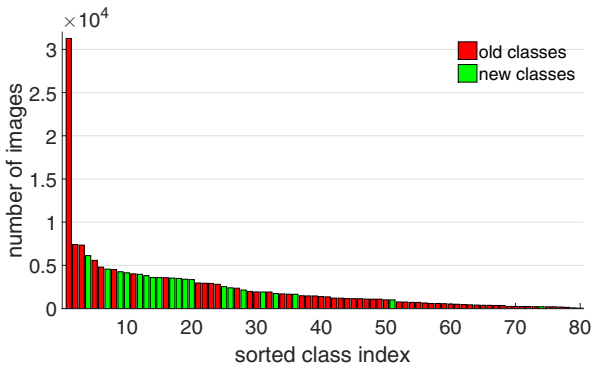

Fig. 2. The data distribution of COCO dataset for the 4-th (final) incremental learning stage. The x-axis represents the sorted category index. The y-axis is the number of available training images for each class. It's worth noting that the old classes have no annotations during training.

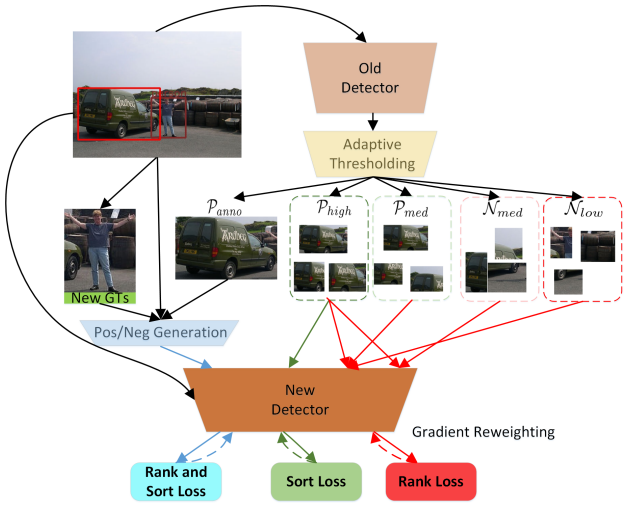

Fig. 3. System overview. Red boxes are pseudo ground truths for old classes; green boxes are ground truths for new classes.

the incremental object detection problem is how to effectively identify and transfer the knowledge learned from previous stages while learning new classes, using only the newest training data.

The contributions of this paper are as follows: (1) We propose an incremental ranking and sorting framework, which is an unified formulation for CIOD that can be applied to various detectors and task settings. (2) We propose an adaptive thresholding method to deal with the problems of pseudo ground truth labelling and positive/negative sampling. (3) We propose a gradient rebalancing strategy to mitigate the problem of imbalanced class distributions for CIOD. (4) We experimentally demonstrate that the BRS model can achieve state-of-the-art performance on standard class-incremental object detection tasks.

2. METHODOLOGY

One-stage detectors perform dense prediction based on a set of anchor boxes or center/corner points, while two-stage detectors perform classification and box regression with a set of region proposals. For CIOD, in every new stage, n_{new} new classes will be added. When testing, the detector need to localize all instances of $n_{old} + n_{new}$ classes observed thus far. Using individual classifier with sigmoid for each class, the final score of a sample with logit s for one class is computed using $\sigma = sigmoid(s)$.

2.1. Incremental Ranking and Sorting

As shown in Fig. 3, the proposed incremental ranking and sorting decomposes the CIOD problem into two tasks: ranking & sorting with ground truths and ranking & sorting transfer. Here ranking aims to rank each positive higher than all negatives, while sorting is designed to sort the logits $\boldsymbol{s}$ of positive samples in descending order w.r.t., continuous ground-truth labels $\boldsymbol{y}$. IoU (Intersection over Union) is used as the label in this study.

Ranking & Sorting with Pseudo Ground Truths (RSP). Given logit s_i and corresponding label $y_i \in [0, 1]$, RS loss [15] is defined as the average of the differences between the predicted $\ell_{\text{RS}}(i)$ and target $\ell^*_{\text{RS}}(i)$ RS errors over positives (i.e. $y_i > 0$):

$$\mathcal{L}_{\text{RS}} := \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \left(\ell_{\text{RS}}(i) - \ell^*_{\text{RS}}(i)\right), \tag{1}$$

where $\ell_{\text{RS}}(i)$ is a summation of the predicted ranking error and current sorting error:

$$\ell_{\text{RS}}(i) := \ell_{\text{R}}(i) + \ell_{\text{S}}(i), \tag{2}$$

where $\ell_{\text{R}}(i)$ and $\ell_{\text{S}}(i)$ are defined in [15]. Let $\mathcal{P}$ and $\mathcal{N}$ denote the sets of positive and negative samples respectively. Then for $i \in \mathcal{N}$, the gradient is computed as:

$$\frac{\partial \mathcal{L}_{\text{RS}}}{\partial s_i} = \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \ell_{\text{R}}(j) p_R(i|j), \tag{3}$$

Gradient for $i \in \mathcal{P}$ is computed as:

$$\frac{\partial \mathcal{L}_{\text{RS}}}{\partial s_i} = \frac{\ell^*_{\text{RS}}(i) - \ell_{\text{RS}}(i) + \sum\limits_{j \in \mathcal{P}} \left(\ell_{\text{S}}(j) - \ell^*_{\text{S}}(j)\right) p_S(i|j)}{|\mathcal{P}|}. \tag{4}$$

Ranking & sorting can be directly used for the training of new classes with RS loss $\mathcal{L}^{new}_{\text{RS}}$ and box regression loss $\mathcal{L}^{new}_{box}$. However, for the old classes, the ground truth labels are not available. Given the old model and a training image x in the new stage, the class-wise logits for old classes can be obtained. We propose to produce a pseudo ground truth set $\mathcal{P}_{anno}$ using samples with logits $s > th_{anno}$, where th_{anno} is a threshold. Then, ranking and sorting for samples of old

classes in the new stage are feasible now with $\mathcal{P}_{anno}$ serving as the substitute of true annotations. We denote the associated losses as $\mathcal{L}_{\text{RS}}^{old}$, $\mathcal{L}_{box}^{old}$, and $\mathcal{L}_{RSP} = \mathcal{L}_{\text{RS}}^{old} + \mathcal{L}_{box}^{old}$.

Ranking & Sorting Transfer (RST). Another method to transfer the knowledge from the old model to the new model, ranking & sorting transfer, is also proposed. Three thresholds, th_{low}, th_{med} and th_{high}, are used to group the anchors/proposals to 4 groups: positive sample set $\mathcal{P}_{high}$, ambiguous positive sample set $\mathcal{P}_{med}$, ambiguous negative sample set $\mathcal{N}_{med}$ and negative sample in set $\mathcal{N}_{low}$, where samples in $\mathcal{P}_{high}$ are with $s > th_{high}$, negative samples $\mathcal{N}_{low}$ are with $s < th_{low}$, samples in $\mathcal{P}_{med}$ are with $th_{med} < s < th_{high}$ and samples in $\mathcal{N}_{med}$ are with $th_{low} < s < th_{med}$. Using logits as labels, sorting is performed on high quality positive samples $\mathcal{P}_{high}$, and ranking is performed between $\{\mathcal{P}_{high}, \mathcal{P}_{med}\}$ and $\mathcal{N}_{low}$, and also between $\mathcal{P}_{high}$ and $\mathcal{N}_{med}$. We denote corresponding loss as $\mathcal{L}_{\text{RS}}^{trans}$. Box regression can also be carried out using old outputs of positive samples now with $\mathcal{L}_{box}^{trans}$. The full loss for RST is defined as: $\mathcal{L}_{RST} = \mathcal{L}_{\text{RS}}^{trans} + \mathcal{L}_{box}^{trans}$.

Adaptive Thresholding. The thresholds for each new class are computed after the training is finished in that stage. Taking all the training data of this class into account, we let N and N_{pos} denote the total number of ground truth instances and positive samples, respectively. Then, all training images are fed to the detector. Two sets, $\{s_i^b\}_{i=1}^{N_{pos}}$ and $\{s_i^a\}_{i=1}^{N}$, are maintained to record the top N_{pos} and N logits of all the output boxes before and after NMS (Non-Maximum Suppression). After sorting the elements in the two sets in descending order, the 4 thresholds can be obtained with $th_{anno} := s_{2N}^a$, $th_{high} := s_{N_{pos}/2}^b$, $th_{med} := s_{3N_{pos}/4}^b$, and $th_{low} := s_{N_{pos}}^b$.

The incremental ranking and sorting loss is defined as:

$$\mathcal{L}_{IRS} = \mathcal{L}_{\text{RS}}^{new} + \mathcal{L}_{\text{RS}}^{old} + \mathcal{L}_{\text{RS}}^{trans}. \tag{5}$$

2.2. Rebalancing Strategy

Although Rank & Sort loss is robust to the imbalance between the positive and negative samples. It may still suffer from class imbalance of data distribution. We propose a weighting strategy to deal with this issue. Inspired by [16], we keep recording the accumulated instance number N_i for each category i at each iteration for every training epoch. Given an instance with positive label i, for another category j, the gradients for the negative sample is weighted using:

$$w_{ij}^b = \begin{cases} 1, & \text{if } N_i \leq N_j \\ \left(\frac{N_j}{N_i}\right)^{\alpha}, & \text{if } N_i > N_j \end{cases} \tag{6}$$

where the exponent α is a hyper-parameter to control the scale. Besides, a difficulty rebalancing strategy is also proposed, which focuses on missorted samples for tail classes. Specifically, a penalty weight is used to strengthen the diminished gradient when the predicted probability σ_j of a positive sample j (with lower label value) is greater than σ_i. The penalty weight w_{ij}^p is computed with another exponent β as:

$$w_{ij}^p = \left(\frac{\sigma_j}{\sigma_i}\right)^{\beta}, \text{ if } \sigma_j > \sigma_i \tag{7}$$

Gradient rebalancing on classes and sample pairs for $\mathcal{L}_{IRS}$ results in the balanced ranking and sorting loss: $\mathcal{L}_{BRS}$.

Table 1. Results (mAP %) on VOC dataset with 5 new classes in every stage, using Faster-RCNN.

	5	10	15	20
ILOD [8]	57.6	55.9	53.7	47.0
RKT [10]	57.6	56.8	56.9	52.9
ILOD Faster [13]	69.6	58.7	51.5	49.3
FILOD [13]	69.6	58.5	53.8	49.7
Ours (RS)	69.8	59.6	57.6	53.2
Ours (BRS)	69.9	**60.4**	**58.3**	**53.6**

Table 2. Results (mAP %) on COCO dataset with 20 new classes in every stage, using Faster-RCNN.

	20	40	60	80
ILOD [8]	22.6	20.3	18.6	16.4
FILOD [13]	29.2	26.3	22.5	19.7
Ours (RST only)	29.5	26.5	23.4	20.5
Ours (RSP only)	29.5	26.7	23.8	20.8
Ours (RS)	29.5	27.1	24.4	21.3
Ours (BRS)	29.6	**27.8**	**24.9**	**21.8**

2.3. Implementation on Detectors

The proposed method can be deployed on various modern object detectors. Feature distillation should be applied to the image feature extractors for fair comparisons with the state-of-the-arts [13, 14]. We let $\mathbf{f} \in \mathbb{R}^{u \times 1}$ denote the flattened feature vector. The feature distillation loss is defined as:

$$\mathcal{L}_{FD} = \frac{1}{2u}||\mathbf{f}_{new} - \mathbf{f}_{old}||_2^2. \tag{8}$$

Faster-RCNN. Faster-RCNN [2] is one of the most popular two-stage object detectors. The RPN can be learned incrementally using knowledge distillation for old classes and regular training for new classes [13]. So the total loss for training Faster-RCNN incrementally is:

$$\mathcal{L}_{FRCN} = \mathcal{L}_{BRS} + \mathcal{L}_{box} + \mathcal{L}_{FD} + \mathcal{L}_{KD}^{RPN}, \tag{9}$$

where $\mathcal{L}_{box}$ is the overall box regression loss.

RetinaNet. The BRS can be directly deployed on the anchors and make dense predictions with RetinaNet [4], which is a single-stage object detector. In addition, feature distillation on FPN is also used. The overall loss is:

$$\mathcal{L}_{RetinaNet} = \mathcal{L}_{BRS} + \mathcal{L}_{box} + \mathcal{L}_{FD}. \tag{10}$$

FCOS. FCOS [5] uses center candidates instead of anchors for dense box prediction, in which IoUs can not be computed. We propose computing a similar metric using the distance

Table 3. Results (mAP %) on COCO dataset using RetinaNet.

	20	40	60	80
KD	30.5	27.3	23.2	20.4
Ours (RS)	30.7	28.9	24.9	22.0
Ours (BRS)	30.8	**29.6**	**25.8**	**22.9**

Table 4. Results (mAP %) on COCO dataset using FCOS.

	20	40	60	80
SID [14]	31.2	28.4	23.9	21.6
Ours (CRS)	31.4	29.3	25.2	22.8
Ours (CBRS)	31.5	**30.7**	**26.4**	**23.6**

between the center candidates and one ground truth central point, d. Let r denote the sampling radius for a feature level. The new continuous label is $y = max(0, 1 - d/r)$. Let's denote this method as CBRS (center guided BRS). With $\mathcal{L}_{center}$ denoting the centerness loss [5], the total loss is:

$$\mathcal{L}_{FCOS} = \mathcal{L}_{CBRS} + \mathcal{L}_{box} + \mathcal{L}_{FD} + \mathcal{L}_{center}. \quad (11)$$

3. EXPERIMENTS

3.1. Datasets

We evaluate our method on the PASCAL VOC 2007 detection benchmark and the COCO 2014 dataset. We use the standard mean average precision (mAP) of IoU=0.5 for VOC 2007, and mAP weighted across different IoUs from 0.5 to 0.95 for evaluation on COCO. Evaluation on the VOC 2007 is done on the test split (train on train and val split), while for COCO, 5k images in minival subset from the validation set are used.

3.2. Implementation Details

We use the standard implementation for ResNet-based Faster RCNN network [17]. ResNet-50 with/without FPN is used as the backbone in different experimental settings, following [14] [13] [10]. We set the training epoches and learning rates of different learning stages following [13]. The parameters α and β are set to 0.7 and 1.5 respectively.

3.3. Results

To compare fairly with the state-of-the-art methods [13] [10], we follow their settings and carry out incremental object detection experiments of 4 stages. We split the 20 classes in the VOC dataset to 4 groups with 5 classes added every stage. Similarly, for the COCO dataset 20 new classes will be added every stage. We compare our method with ILOD [8], RKT [10], ILOD Faster [13], FILOD [13] and SID [14]. As shown in Table 1, compared with ILOD and RKT, our method performs much better in terms of mAP using Faster-RCNN. However, these two methods do not use modern RPN to generate proposals. So for fair comparison, we take ILOD Faster and FILOD as the baselines. On VOC dataset our method outperforms them by 4.3% and 3.9% mAP after learning all 4 stages. Results on COCO are shown in Table 2. Significant mAP leap is achieved compared to FILOD along the learning stage. The final improvement over ILOD and FILOD are 5.4% and 2.1% respectively. For 6-stage incremental learning, as shown in Table 5, our method also outperforms state-of-the-art methods by a margin. We also carry out ablation study to analyze the two components of our method: incremental ranking and sorting, and gradient rebalancing. Table 1 shows that with incremental ranking and sorting only our method already outperforms FILOD by 3.5% mAP. This verifies the effectiveness of incremental ranking and sorting. Rebalancing brings another 0.4% mAP gain. The result is consistent on COCO, where 1.6% and 0.5% absolute improvements on mAP are achieved. The initial mAP gains of our method are small, which verifies that our method is effective for CIOD, rather than fully supervised learning. It's worth noting that using RSP or RST only the proposed method outperforms FILOD by 1.1% or 0.8% final mAP, respectively, indicating that either RSP or RST is effective.

Table 5. Results (mAP %) on VOC dataset under the 15+1+1+1+1+1 setting.

	15	16	17	18	19	20
FILOD [13]	73.1	70.1	68.3	65.9	63.7	61.3
Ours	73.2	**72.2**	**70.1**	**68.6**	**65.8**	**63.9**
SID [14]	73.7	68.0	63.0	57.3	53.2	48.9
Ours	73.7	**69.3**	**65.4**	**59.8**	**55.9**	**51.7**

We further carry out experiments to evaluate our method on two one-stage object detectors: anchor-based RetinaNet and anchor-free method FCOS. For these two detectors experiments are performed on COCO dataset. As shown in Table 3 and 4, the initial mAPs are boosted from faster-RCNN because of the use of FPN. Specifically, based on RetinaNet, our method outperforms knowledge distillation (KD) by 2.5% mAP after learning all 4 stages. When using FCOS, our method achieves 2.0% mAP gain over a strong baseline, SID. Ablation studies shows that, based on RetinaNet, 1.6% and 0.9% mAP gain can be achieved by incremental ranking and sorting, and rebalancing respectively. While the improvements on mAP when using FCOS are 1.2% and 0.8%. Rebalancing is more effective for one-stage detectors and large dataset with severe data imbalance.

4. CONCLUSION

In this study, we propose balanced ranking and sorting, to tackle the catastrophic forgetting and data imbalance problems in class incremental object detection. We propose to incrementally rank and sort samples to learn new classes while transfer the knowledge from the old model to the new one, with unified loss design. Gradient rebalancing is performed using two effective strategies to deal with data imbalance. We implement BRS on several popular object detectors. Extensive experiments on PASCAL VOC and COCO datasets show that our method outperforms the state-of-the-arts.

5. REFERENCES

[1] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

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

[5] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.

[8] K. Shmelkov, C. Schmid, and K. Alahari, "Incremental learning of object detectors without catastrophic forgetting," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3400–3409.

[9] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[10] K. Ramakrishnan, R. Panda, Q. Fan, J. Henning, A. Oliva, and R. Feris, "Relationship matters: Relation guided knowledge transfer for incremental learning of object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020.

[11] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*, 2014, pp. 391–405.

[12] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 328–335.

[13] C. Peng, K. Zhao, and B. C. Lovell, "Faster ilod: Incremental learning for object detectors based on faster rcnn," *Pattern recognition letters*, vol. 140, pp. 109–115, 2020.

[14] C. Peng, K. Zhao, S. Maksoud, M. Li, and B. C. Lovell, "Sid: Incremental learning for anchor-free object detection via selective and inter-related distillation," *Computer vision and image understanding*, p. 103229, 2021.

[15] K. Oksuz, B. C. Cam, E. Akbas, and S. Kalkan, "Rank & sort loss for object detection and instance segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2021.

[16] J. Wang, W. Zhang, Y. Zang, Y. Cao, J. Pang, T. Gong, K. Chen, Z. Liu, C. C. Loy, and D. Lin, "Seesaw loss for long-tailed instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9695–9704.

[17] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1476–1481, 2016.