Reconcile Prediction Consistency for Balanced Object Detection

Keyang Wang, Lei Zhang$^{(\boxtimes)}$

Learning Intelligence & Vision Essential (LiVE) Group
School of Microelectronics and Communication Engineering, Chongqing University, China

{wangkeyang, leizhang}@cqu.edu.cn

Abstract

Classification and regression are two pillars of object detectors. In most CNN-based detectors, these two pillars are optimized independently. Without direct interactions between them, the classification loss and the regression loss can not be optimized synchronously toward the optimal direction in the training phase. This clearly leads to lots of inconsistent predictions with high classification score but low localization accuracy or low classification score but high localization accuracy in the inference phase, especially for the objects of irregular shape and occlusion, which severely hurts the detection performance of existing detectors after N-MS. To reconcile prediction consistency for balanced object detection, we propose a Harmonic loss to harmonize the optimization of classification branch and localization branch. The Harmonic loss enables these two branches to supervise and promote each other during training, thereby producing consistent predictions with high co-occurrence of top classification and localization in the inference phase. Furthermore, in order to prevent the localization loss from being dominated by outliers during training phase, a Harmonic IoU loss is proposed to harmonize the weight of the localization loss of different IoU-level samples. Comprehensive experiments on benchmarks PASCAL VOC and MS COCO demonstrate the generality and effectiveness of our model for facilitating existing object detectors to state-of-the-art accuracy.

1. Introduction

Object detection serves as a prerequisite for a broad set of downstream vision applications, such as person re-identification [3], instance segmentation [9] and action recognition [8]. With the foundation of deep neural networks, object detection has achieved significant advances in recent years. On the whole, all the detectors that use deep CNN can be divided into two categories: (1) The multi-stage approaches, including [7, 22, 15, 1, 14, 30]. For the multi-stage object detectors, multi-stage classification and localization are applied sequentially, which make these models more powerful
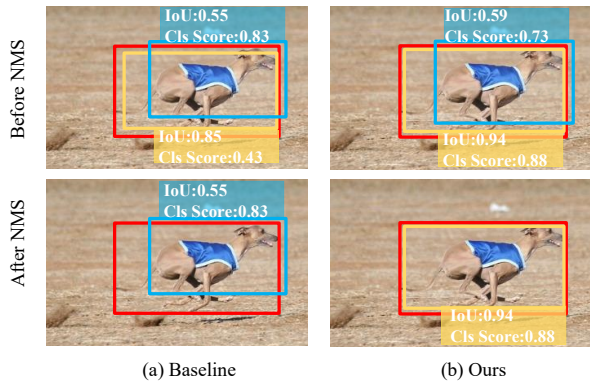

Figure 1. Irregular (e.g., tail) and occluded (e.g., background) object detection outputs of the baseline detector and our detector trained with Harmonic loss. The red bounding boxes denote the ground-truth. The baseline detector (a) produces inconsistent bounding boxes with high classification score but low localization IoU (the blue bbox), or low classification score but high localization IoU (the yellow bbox), which will hurt the detection accuracy after NMS (*i.e.*, suboptimal blue bbox is selected). Our detector (b) produces bounding boxes with high co-occurrence of top classification and localization by harmonizing the optimization of these two branches, leading to better detection results after NMS.

on classification and localization tasks. So these approaches have achieved top performance on benchmark datasets. (2) The one-stage approaches, including [19, 21, 16, 29, 26]. The one-stage approaches apply object classifiers and regressors in a dense manner without object-based pruning. The main advantage of the one-stage detectors is their inference efficiency, but the detection accuracy is usually inferior to the two-stage approaches.

Modern CNN-based object detectors divide the object detection task into two branches, classification branch and regression branch, which are trained with independent objective functions without knowing each other explicitly. Specifically, a classification loss (e.g., cross-entropy loss [7]) is adopted to optimize the classifier and a distance loss (e.g., smoothL1 loss [7]) is used to optimize the regressor. However, since the correlation between classification task and localization task is ignored in existing detectors, the opti-

mization of the classifier and the regressor often can not be synchronous. Specifically, the classification loss drives the model to learn as high classification score as possible for all the positive examples regardless of the localization accuracy, and the regression loss enables the model to output accurate regression offset for each anchor without taking into account the classification score during training. As a result, this leads to serious prediction inconsistency between classification and regression. Specifically, the detector outputs lots of inconsistent detection results that have high classification score but low IoU or low classification score but high IoU in the inference phase. As shown in Fig. 1 (a), we show a representative example of inharmonious bounding boxes detected by the baseline. Two inharmonious candidates (a blue bbox and a yellow bbox) are produced by the baseline. Clearly after the NMS procedure, the yellow one with high IoU but low classification score will be suppressed by the less accurate blue one. This means that, for this instance, the detector will end up with a clearly suboptimal detection result (*i.e.*, blue bbox), but not the optimal one (*i.e.*, yellow bbox). Since classification loss and regression loss are independently optimized for each positive sample throughout the training phase, such inconsistency is very common in the testing phase, especially for the objects of irregular shape and occlusion, which will hurt the overall detection performance. However, the prediction inconsistency between tasks is paid less attention in object detection community.

In order to reconcile prediction consistency, we propose a mutually supervised loss, called Harmonic loss, to jointly optimize the classification and regression branches toward the optimal direction. In the training phase, the classification branch will continuously supervise the optimization status of the regression branch through a harmonic factor related to the classification loss, and vice versa. By plugging and playing the Harmonic loss into existing detectors, the network will become training-aware and reduce the training imbalance across tasks by mutual interaction. That is, the easily-trained branch will learn to dynamically facilitate the learning of the other branch towards a progressively harmonic state. As a result, the detectors produce the bounding boxes with high co-occurrence of top classification and localization in the testing, thereby alleviating the prediction inconsistency between these two tasks, as is shown in Fig. 1 (b).

Furthermore, during the training phase of the detectors, only positive samples are used to optimize the regression branch, but the IoU distribution of all the positive samples is seriously imbalanced (Please see Fig.3 (a) in the method section). Generally, the number of samples at low IoU levels is significantly larger than that at high IoU levels. This will cause that the localization losses of existing detectors, especially one-stage detectors, are dominated by outliers (low IoU levels) during training phase, which enables the detection model to be biased toward the outliers. In order to

alleviate the bias, motivated by focal loss [16], we propose a simple but effective Harmonic IoU (HIoU) loss to dynamically harmonize the contribution of each kind of examples in training phase. We up-weight the weights of localization loss of examples with high IoU while suppressing the weights of examples with low IoU by a dynamic scaling factor. By embedding this HIoU loss into our proposed Harmonic loss, the contribution of each level of samples is balanced and the training is more effective and stable.

By embedding the HIoU loss into our proposed Harmonic loss, a Harmonic detection (HarmonicDet) loss is formulated. We plug and play our HarmonicDet loss into both two-stage and single-stage detection frameworks on the PASCAL VOC [6] and MS COCO [17] benchmarks. The detection results demonstrate the effectiveness and generality of our approach in improving existing detectors. In summary, this paper makes three contributions:

1. We propose a Harmonic loss to jointly optimize the classification and regression branch toward the optimal direction in the training phase, thereby reconciling prediction consistency between classification and localization in the testing phase.
2. We introduce a Harmonic IoU (HIoU) loss to harmonize the weight of the localization loss of different IoU-level samples, which prevents the localization loss from being dominated by outliers and ensures accurate bounding box regression ability of the detector.
3. The proposed losses can be easily plugged and played into different state-of-the-art detection algorithms, achieving significant performance gains on the PASCAL VOC dataset and MS COCO dataset.

2. Related Work

Architecture design for object detection. Benefit from the Deep ConvNets, CNN has achieved a great success in the object detection field. All the CNN-based detectors can be roughly divided into two categories, *i.e.,* two-stage detectors and one-stage detectors. The two-stage detectors consist of two parts. The first part is responsible for generating a set of candidate object proposals, e.g., Selective Search [25], EdgeBoxes [32]. The second part determines the accurate object regions and the corresponding class labels according to the candidate object proposals. Its descendants (e.g., Fast R-CNN [7], Faster R-CNN [22], R-FCN [4], FPN [15], Mask RCNN [9]) achieve dominant performance on several challenging datasets. In contrast, one-stage approaches like SSD [19] directly predict the class scores and location offsets of the default bounding boxes, without the region proposal step. After that, more advanced one-stage detectors (e.g., RetinaNet [16], RefineDet [29], RFB [18]) are proposed to achieve high detection accuracy. In order to solve the problem that the detection result is too sensitive to the size of the anchor and simultaneously avoid the complex IoU compu-

tation between anchor boxes and ground-truth boxes during training, some anchor-free detectors are proposed, including CornerNet [12], FCOS [24], CenterNet [5], FASF [31].

Loss design for accurate object detection. In order to strengthen the correlation between classification and localization tasks and get accurate bounding boxes, some detectors have also made some representative explorations. IoU-Net [11] integrates the IoU predictor into existing object detectors to predict the IoU between each detected bounding box and the matched ground-truth. Simultaneously, an IoU-guided NMS is proposed to resolve the misalignment between classification confidence and localization accuracy. PISA [2] uses the classification probabilities to reweight the contribution of each example to the regression loss, such that the regression branch can be strengthened for improving detection performance. In order to balance the contribution of each kind of samples to the localization branch, some sample-adaptive localization losses are studied. GHM [13] analyzes the sample imbalance in one-stage detectors in term of gradient norm distribution and propose a gradient aware localization loss GHM-R. Libra R-CNN [20] claims that the small gradients produced by the easy samples may be neglected comparing to the large gradients produced by the outliers and a balanced $\mathcal{L}_1$ loss is proposed to increase the gradient of easy samples and keep the outliers unchanged.

3. Proposed Approach

In this section, we will firstly introduce the Harmonic loss in section 3.1. In particular, the in-depth insight on how it can harmonize the optimization from the loss and its gradient is presented. Then we further elaborate the proposed Harmonic IoU (HIoU) loss in section 3.2.

3.1. Harmonic Loss

We firstly revisit the standard loss of object detector, as shown in Eq.(1).

$$\mathcal{L}_{Det} = \frac{1}{N} \left(\sum_{i \in Pos} (CE(p_i, y_i) + \mathcal{L}(d_i, \hat{d}_i)) + \sum_{j \in Neg} CE(p_j, y_j) \right)$$
(1)

where *N* and *M* is the number of *Positive* and *Negative* samples, resp. And p_i and y_i denotes the predicted classification score and the corresponding ground truth class, resp. d_i and $\hat{d}_i$ denotes the output regression offset and target offset, resp. CE(·) is the cross entropy loss and $\mathcal{L}(\cdot)$ is the commonly used Smooth L1 loss,

We can clearly find that for each positive sample, the classification and localization branches are trained with independent objective functions without knowing each other explicitly, which causes serious prediction inconsistency between classification and regression tasks. For example, for classification branch, the CE loss drives the model to learn as high classification score as possible for all samples regardless of the localization accuracy during training. As a result,

the predicted classification scores are independent of the localization accuracy and there must be many inconsistent detections that have high classification score but low IoU or low classification score but high IoU, which deteriorates the overall detection performance after NMS.

In order to avoid these inconsistent detections, we propose a Harmonic loss to harmonize the training of classification and regression branches and strengthen the correlation between classification score and localization accuracy. Since only positive samples need to be regressed, we only apply Harmonic loss on positive samples. So given a positive sample x_i, the Harmonic loss can be defined as follows:

$$\mathcal{L}_{Har}^i = (1 + \beta_r) CE(p_i, y_i) + (1 + \beta_c) \mathcal{L}(d_i, \hat{d}_i)$$
(2)

where β_r, β_c are two key but dynamic harmonic factors in our method, which can be defined as follows:

$$\beta_r = e^{-\mathcal{L}(d_i, \hat{d}_i)} \quad , \quad \beta_c = e^{-CE(p_i, y_i)}$$
(3)

From Eq.(2) and Eq.(3), we can clearly find that there are two mutually interacted parts in our Harmonic loss. We use a mutual supervision mechanism to harmonize the optimization of classification and regression branch. For classification task, a regression loss derived harmonic factor $(1 + \beta_r)$ is assigned to dynamically supervise the optimization of the classification branch. In other words, a regression-aware classification loss is devised to optimize the classification branch. Similarly, for regression task, a dynamic factor $(1 + \beta_c)$ related to the classification loss is used to supervise the optimization of regression loss, which means that the regression loss is classification-aware. In the following, in-depth insight and detailed analysis on why the Harmonic loss can harmonize the classification and regression branch are elaborated from the loss function and the gradient. And due to space limitation, we give the convergence analysis of our harmonic loss in the **supplementary material**.

Effectiveness analysis from the loss function. Given an object x_i, and suppose the predicted class score w.r.t. its ground truth class to be p_i and the location accuracy w.r.t. the ground truth bounding box to be IoU_i. We may face four different situations during the optimization process: **(1)** *classification is good* ($p_i \uparrow$) *and regression is good* ($IoU_i \uparrow$). This means that the sample x_i is the prime sample for learning high classification scores and accurate localization. So we simultaneously assign a large harmonic factor β_r for classification loss according to the state of regression, and a large harmonic factor β_c for regression loss according to the state of classification. **(2)** *classification is good* ($p_i \uparrow$) *but regression is bad* ($IoU_i \downarrow$). This is one of the inharmonious training situations that we need to solve. Since the regression task for this positive sample is poor, we assign a large harmonic factor β_c for regression loss according to the the state of classification and a small harmonic weight β_r for classification loss according to the state of regression. Intuitively,

the network will reinforce attention to primarily optimize the regression branch. In addition, since the dynamic harmonic factor β_c is derived by the classification loss, this means that the regressor can always perceive the classification status. In other words, the classification loss will also supervise the optimization of the regression branch. This is also true for classification branch. As a result, these two branches will supervise and promote each other, and optimize simultaneously toward the optimal direction. The disharmony between classification and regression is therefore alleviated. **(3)** *classification is bad* ($p_i \downarrow$) *but regression is good* ($IoU_i \uparrow$). This is another inharmonious training situation. Contrary to situation (2), a small harmonic weighting factor β_c and a large harmonic weighting factor β_r is assigned for regression loss and classification loss, respectively. As a result, the network primarily optimize the classification branch and alleviate the disharmony between the two tasks. From situation (2) and (3), we see that in the Harmonic loss the easily-trained branch is always enabled to dynamically facilitate the learning of the other branch. **(4)** *classification is bad* ($p_i \downarrow$) *and regression is bad* ($IoU_i \downarrow$). This means that the positive sample x_i is a hard positive sample. We claim that those hard positive samples are not the prime samples and have little effect on the optimization of the detectors. So we simultaneously decrease the weights β_r, β_c of the classification loss and localization loss, resp. By applying the Harmonic loss to optimize the existing detectors during training, these detectors can naturally harmonize the classification score and localization accuracy in the testing phase.

Effectiveness analysis on classification task from the gradient. We have described in detail how to apply the Harmonic loss to achieve the interaction and harmony between classification task and regression task from the aspect of loss function. In the following, we will further discuss how our Harmonic loss harmonizes the training of the two tasks from the aspect of gradient. For classification task, we calculate the partial derivative of Eq.(2) with respect to the predicted score p_i of the corresponding ground truth class as:

$$\frac{\partial \mathcal{L}_{Har}^i}{\partial p_i} = (1 + \beta_r)\frac{\partial CE(p_i, y_i)}{\partial p_i} + \mathcal{L}(d_i, \hat{d}_i)\frac{\partial \beta_c}{\partial p_i} \quad (4)$$

where $CE(\cdot)$ is the cross-entropy loss and y_i is the one-hot label, and there is

$$CE(p_i, y_i) = -y_i \log(p_i) = -\log(p_i),$$
$$\beta_c = e^{-CE(p_i, y_i)} = e^{y_i \log(p_i)} = p_i \quad (5)$$

By substituting Eq.(5) into Eq.(4), we can have

$$\frac{\partial \mathcal{L}_{Har}^i}{\partial p_i} = -(1 + \beta_r)\frac{1}{p_i} + \mathcal{L}(d_i, \hat{d}_i) \quad (6)$$

By substituting $\beta_r = e^{-\mathcal{L}(d_i, \hat{d}_i)}$ into Eq.(6), we can have

$$\frac{\partial \mathcal{L}_{Har}^i}{\partial p_i} = \mathcal{L}(d_i, \hat{d}_i) - \frac{(1 + e^{-\mathcal{L}(d_i, \hat{d}_i)})}{p_i} \quad (7)$$
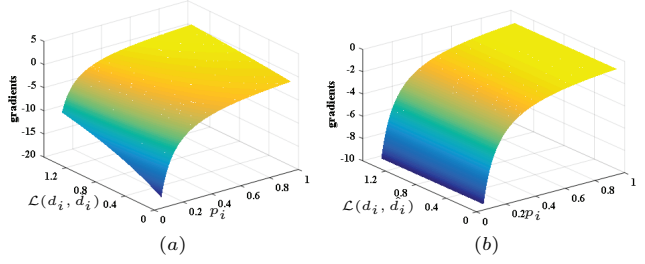

Figure 2. Visualization of the gradients of detection losses with respect to the predicted score of the corresponding ground truth class p_i. (a) is the gradient of our Harmonic loss with respect to p_i, $\frac{\partial \mathcal{L}_{Har}^i}{\partial p_i}$. (b) is the gradient of the standard detection loss (cross-entropy loss plus smooth L1 loss) with respect to p_i.

We can clearly find that the partial derivative of our Harmonic loss to the classification probability p_i is correlated the regression loss, which means that the classification branch will be supervised by the regression loss during the optimization process. In order to further illustrate the effect of the localization loss to the gradient in Eq.(7), we visualize the gradients of the standard detection loss and the proposed Harmonic loss with respect to p_i, respectively, in Fig. 2. For the standard detection loss, the gradient with respect to p_i does not change with different location loss values as Fig. 2 (b) shows, which means the optimization of classification is absolutely independent of the regression task. But for our Harmonic loss, the gradient is a function simultaneously determined by the two variables p_i and $\mathcal{L}(d_i, \hat{d}_i)$. In other words, for each positive sample, the localization loss $\mathcal{L}(d_i, \hat{d}_i)$ will supervise the optimization of the classification branch during training phase of the Harmonic loss. Specifically, in Eq.(7), there is a proportional correlation between $\mathcal{L}(d_i, \hat{d}_i)$ and the loss gradient w.r.t. p_i. But since this gradient value is always negative during the training process, the absolute value of this gradient decreases as the regression loss increases, as is shown in Fig. 2 (a). This means that the gradient will suppress the classification scores of samples with low regression quality.

Effectiveness analysis on regression task from the gradient. The mechanism of our Harmonic loss for regression task is similar to classification task. For regression task, the gradient of Harmonic loss with respect to the predicted regression offset $(d_i - \hat{d}_i)$ is also supervised by the classification score p_i. Due to space limitation, we give this derivation process in the **supplementary material**.

Task-Contrastive Loss. In order to further reduce the gap between the two tasks, we introduce a simple Task-Contrastive (TC) loss, as defined in Eq.(8), to our Harmonic loss. We directly reduce the contrastive distance between the predicted score p_i and the localization accuracy IoU_i for the sample x_i w.r.t. its ground truth class and bounding box. A margin is adopted because we think that the gap between classification and regression can be ignored when

the distance between p_i and IoU_i is smaller than the margin. Additionally, we assign an information entropy-guided weight factor $\frac{1}{1+\beta_e}$ for each sample. Specifically, when the entropy is large, it means that the classification uncertainty is large, so a small weight, $\frac{1}{1+\beta_e}$, is assigned, and vice versa.

$$\mathcal{L}_{TC}^i = \frac{1}{1+\beta_e} \max(0, |p_i - IoU_i| - margin)$$

$$\beta_e = e^{-\sum_{k=1}^{C} p_k \log(p_k)}$$

(8)

where C is the number of classes, and the *margin* is set as 0.2 in our experiments.

3.2. Harmonic IoU Loss

The IoU based localization loss [28] has proved to be a suitable choice to obtain the optimal IoU metric, which is written as:

$$\mathcal{L}_{IoU}^i = 1 - IoU_i \qquad (9)$$

However, the typical IoU loss treats all the positive samples equally and neglects a fact that IoU distribution is seriously imbalanced. As shown in Fig. 3 (a), we visualize the IoU distribution of the 100k positive samples. We can find that the number of samples at low IoU levels is much larger than that at high IoU levels. Needless to say, the hard samples at low IoUs dominate the gradients of the localization loss in the training phase, which enables the detection model to be biased toward the hard samples (low IoU levels). As shown in Fig. 3 (b), we visualize the average localization improvement of proposals from different IoU intervals after refinement. We can find that, with the increase of IoU, the gain of refinement gets smaller and the performance even gets worse at high IoU levels (IoU=0.8, 0.9). This is because the regressor always outputs large offset in the testing phase as it is biased towards hard samples. In order to alleviate the bias, motivated by focal loss [16], we introduce a Harmonic IoU (HIoU) loss, defined as:

$$\mathcal{L}_{HIoU}^i = (1 + IoU_i)^\gamma (1 - IoU_i) \qquad (10)$$

γ is the tunable focusing parameter. A dynamic factor $(1 + IoU_i)^\gamma$ that can automatically reflect the localization accuracy is used to harmonize the weight of the localization loss of different IoU-level samples. Specifically, we use the factor to increase the weights of localization loss of examples with high IoU while suppressing the weights of examples with low IoU. In the end, the contribution of each loss of examples is balanced and the bias of regression can be effectively alleviated. As shown in Fig. 3 (b), the model trained with our Harmonic IoU loss can prevent the localization loss from being dominated by outliers during training phase and ensure accurate bounding box regression ability of the whole detector. Let us take the samples with IoU greater than 0.8 before refinement as an example. The original localization loss causes negative regression, but our Harmonic IoU loss can effectively avoid this situation. It should be noted that in order to ensure that the Harmonic IoU
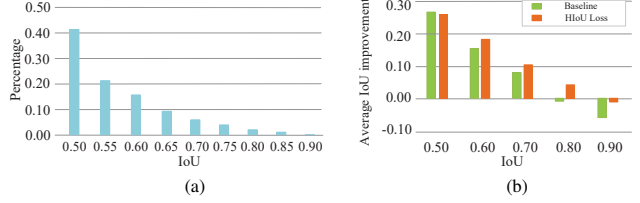

Figure 3. Two histograms of the positive samples. (a) is the IoU distribution of the 100k positive samples before refinement. (b) is the average localization improvement of proposals from different IoU intervals after refinement.

loss is always monotonic, the value of γ needs to satisfy $\gamma \leq 1$, we give the derivation in the **supplementary materials**.

Full Localization Loss. It should be noted that the SmoothL1 loss can directly optimize the distance between central points of the predicted boxes and GT boxes, thereby accelerating the convergence speed of training. So we do not replace the SmoothL1 loss in our Harmonic loss with HIoU loss, but add the HIoU loss as a part of the localization loss. So the full localization loss can be defined as follows:

$$\mathcal{L}_{loc}^i = \mathcal{L}(d_i, \hat{d}_i) + \alpha \mathcal{L}_{HIoU}^i \qquad (11)$$

where α is a trade-off parameter.

3.3. Overall Detection Loss

By replacing localization loss $\mathcal{L}(d_i, \hat{d}_i)$ in Eq. (2) with the full localization loss in Eq. (11) and embedding the TC loss as defined in Eq. (8), the overall **Harmonic Det**ection loss for positive sample x_i can be formulated as follows:

$$\mathcal{L}_{HarDet}^i = (1 + \beta_r)CE(p_i, y_i) + (1 + \beta_c)\mathcal{L}_{loc}^i + \mathcal{L}_{TC}^i \quad (12)$$

where $\beta_r = e^{-\mathcal{L}_{loc}^i}, \beta_c = e^{-CE(p_i, y_i)}$ are the two key dynamic harmonic factor derived in Eq.(3). By further adding the classification loss of negative samples, we can get the whole objective function as follows:

$$\mathcal{L}_{HarDet} = \frac{1}{N} \left(\sum_{i \in Pos}^{N} \mathcal{L}_{HarDet}^i + \sum_{j \in Neg}^{M} CE(p_j, y_j) \right) \quad (13)$$

4. Experiments
4.1. Experimental Setting

Datasets and Baselines. We integrate the proposed loss function into six popular one-stage and two-stage detectors including SSD, DSSD, RefineDet, RetinaNet, Faster R-CNN and Mask R-CNN, and conduct experiments on two benchmarks, *i.e.*, PASCAL VOC and MS COCO.

Backbones. ResNet-50 [10], ResNet-101 [10], ResNeXt-101 [27], pretrained on the standard classification task [23], are used as our backbone networks. Specifically, ResNet-50 and ResNet-101 are mainly used in the experiments of PASCAL VOC dataset. ResNet-50, ResNet-101 and ResNeXt-101 are all used in the experiments of MS COCO dataset.

Table 1. Main results on PASCAL VOC 2007 `test` set. Four different detectors are adopted in our experiments.

Method	Backbone	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
two-stage detectors:							
Faster R-CNN	ResNet-50-FPN	51.7	80.9	76.0	65.1	43.8	11.1
Faster R-CNN	ResNet-101-FPN	54.1	81.8	77.4	67.8	48.5	15.0
Faster R-CNN w/HarmonicDet Loss	ResNet-50-FPN	**54.2**(+2.5)	81.7	77.5	67.3	48.2	14.7
Faster R-CNN w/HarmonicDet Loss	ResNet-101-FPN	**56.5**(+2.4)	82.6	78.7	70.1	52.5	18.2
one-stage detectors:							
SSD300	ResNet-50-FPN	51.0	78.9	73.4	63.3	44.1	14.6
SSD512	ResNet-50-FPN	53.3	81.5	76.8	66.6	47.3	15.3
DSSD320	ResNet-50	52.3	79.8	75.0	64.5	46.2	16.2
DSSD512	ResNet-50	54.2	81.8	77.7	67.4	50.0	15.7
RefineDet320	ResNet-50	53.4	80.2	76.0	65.5	48.0	17.5
RefineDet512	ResNet-50	56.3	82.4	79.3	69.8	52.7	19.6
SSD300 w/HarmonicDet Loss	ResNet-50-FPN	**54.5** (+3.5)	79.7	74.9	65.8	50.2	21.2
SSD512 w/HarmonicDet Loss	ResNet-50-FPN	**56.7** (+3.4)	81.8	78.0	68.9	52.7	22.0
DSSD320 w/HarmonicDet Loss	ResNet-50	**55.7** (+3.4)	80.0	75.7	66.9	52.1	23.7
DSSD512 w/HarmonicDet Loss	ResNet-50	**57.3** (+3.1)	82.4	78.6	69.0	53.4	23.1
RefineDet320 w/HarmonicDet Loss	ResNet-50	**56.9** (+3.5)	80.2	76.2	67.5	53.2	27.2
RefineDet512 w/HarmonicDet Loss	ResNet-50	**60.4** (+4.1)	83.0	79.7	72.0	58.0	29.7

4.2. Experiments on PASCAL VOC

For PASCAL VOC dataset, three one-stage detectors (SSD, DSSD, RefineDet) and one two-stage detector (Faster R-CNN) are adopted as the detectors in our experiments. All models are trained on the union of the VOC 2007 `trainval` and VOC 2012 `trainval` datasets, and tested on the VOC 2007 `test` set. In order to prove that our model can regress more accurate bounding boxes, a stricter COCO-style Average Precision (averaged AP at IoUs from 0.5 to 0.9) metric is adopted on the PASCAL VOC dataset.

From Table 1, we can see that our HarmonicDet loss achieves consistently significant AP improvements on both the one-stage and two-stage detectors, indicating its effectiveness and generality. For two-stage detectors, our HarmonicDet loss can improve the AP by 2.5% compared with the baselines Faster R-CNN with ResNet-50-FPN backbone. Even with a deeper backbone like ResNet-101-FPN, similar improvements are observed. For one-stage detectors, our HarmonicDet loss improves SSD, DSSD and RefineDet by 3.5%, 3.4% and 3.5%, respectively, when we adopt the small input size. Even with a larger input size (512×512), our HarmonicDet loss can also improve SSD, DSSD and RefineDet by 3.4%, 3.1% and 4.1%, respectively. Especially, the improvement for AP at higher IoU threshold (0.8, 0.9) is very significant. When the IoU threshold is 0.8, our HarmonicDet loss can improve the AP by $5\% \sim 6\%$. When the IoU threshold is 0.9, our HarmonicDet loss can even improve the AP by 10%. This proves that the detectors trained with our HarmonicDet loss can regress more accurate bboxes.

4.3. Experiments on MS COCO

For MS COCO dataset, three one-stage detectors (SSD, RefineDet, RetinaNet) and two two-stage detector(Faster R-CNN, Mask R-CNN) are adopted as the detectors in our experiments. The results on MS COCO dataset are shown in Table 2, our HarmonicDet loss also achieves consistently

significant AP improvements on both the one-stage and two-stage detectors. Specifically, on SSD300 and SSD512, the gain is 1.9% and 1.7%, respectively, with the ResNet-50-FPN backbone. On RefineDet320 and RefineDet512, the gain is 1.4% and 2.4%, respectively, with the ResNet-101 backbone. For the RetinaNet, our approach can improve the AP by 1.5% with ResNet-50-FPN backbone. Even with a strong backbone like ResNeXt-101-32x8d, our model also outperforms the baseline by 1.4% AP. For two-stage detector, HarmonicDet loss can improve the AP of Faster R-CNN, Mask R-CNN by 1.8%, 1.6%, respectively, with ResNet-50-FPN backbone. When we adopt the ResNeXt-101-FPN as the backbone, our method can achieve 46.9% AP, which is better than some state-of-the-art detectors, such as Cascade R-CNN (42.8%), FCOS (42.7%) and FASF (42.9%).

4.4. Analysis of Prediction Consistency

It is clear that we reconcile the training of classification and regression by designing two dynamic harmonic factors in the Harmonic loss. However, *are these two branches optimized harmonically in the actual training process*? In the following, we will give the detailed analysis of prediction consistency during the training phase from two aspects: Firstly, as shown in Fig. 4 (a), we visualize the dynamic distribution of the two harmonic factors (i.e., $1+\beta_r$, $1+\beta_c$) during the optimization process. Obviously, in the early stage of training, the gap between the two harmonic factors is relatively large because the inconsistency of the two branches is serious. But with the advancement of optimization, the gap between the two harmonic factors gradually decreases and remains stable at the end of the training, which means that the two branches are progressively in harmony. In addition, in order to quantitatively analyze the prediction inconsistency between classification and regression during the optimization, we propose a simple metric named Average Inconsistency Coefficient (AIC), which is defined as follows:

$$AIC = \sum_{i \in Pos}^{N} |P_i - IoU_i| \tag{14}$$

Table 2. Main Results of our experiments on MS COCO `test-dev` set. The ‡ means multi-scale testing

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
state-of-the-art detectors:							
FCOS [24]	ResNeXt-101-FPN	42.7	62.2	46.1	26.0	45.6	52.6
FASF [31]	ResNeXt-101-FPN	42.9	63.8	46.3	26.6	46.2	52.7
Cascade R-CNN[1]	ResNet-101-FPN	42.8	62.1	46.3	23.7	45.5	55.2
one-stage detectors:							
SSD300	ResNet-50-FPN	29.5	49.0	30.5	11.0	32.6	47.0
SSD512	ResNet-50-FPN	34.4	55.0	36.5	16.3	39.6	50.1
RefineDet320	ResNet-101	32.0	51.4	34.2	10.5	34.7	50.4
RefineDet512	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4
RetinaNet	ResNet-50-FPN	36.1	56.0	38.3	19.8	38.9	45.0
RetinaNet	ResNeXt-101-FPN	39.8	59.9	42.9	22.5	43.1	50.9
SSD300 w/HarmonicDet Loss	ResNet-50-FPN	**31.4**(+1.9)	49.3	32.9	11.1	33.3	49.3
SSD512 w/HarmonicDet Loss	ResNet-50-FPN	**36.1**(+1.7)	55.2	38.3	16.8	40.6	51.2
RefineDet320 w/HarmonicDet Loss	ResNet-101	**33.4**(+1.4)	51.6	35.4	11.1	35.2	52.3
RefineDet512 w/HarmonicDet Loss	ResNet-101	**38.8** (+2.4)	56.5	41.8	17.8	44.4	54.8
RetinaNet w/HarmonicDet Loss	ResNet-50-FPN	**37.6**(+1.5)	56.0	40.1	20.0	39.6	48.7
RetinaNet w/HarmonicDet Loss	ResNeXt-101-FPN	**41.2**(+1.4)	60.3	44.2	23.2	43.8	52.4
two-stage detectors:							
Faster R-CNN	ResNet-50-FPN	37.5	59.0	41.0	21.4	40.1	49.5
Faster R-CNN	ResNeXt-101-FPN	41.4	63.6	45.3	25.4	45.5	52.3
Mask R-CNN	ResNet-50-FPN	38.5	59.5	41.9	21.9	40.8	50.1
Mask R-CNN	ResNeXt-101-FPN	42.3	64.1	46.2	26.1	45.9	54.3
Faster R-CNN w/HarmonicDet Loss	ResNet-50-FPN	**39.3**(+1.8)	59.8	42.7	22.0	41.4	50.0
Faster R-CNN w/HarmonicDet Loss	ResNeXt-101-FPN	**43.0**(+1.6)	64.0	47.0	26.0	46.1	53.6
Mask R-CNN w/HarmonicDet Loss	ResNet-50-FPN	**40.1**(+1.6)	60.5	43.7	22.8	42.3	50.7
Mask R-CNN w/HarmonicDet Loss	ResNeXt-101-FPN	**44.0**(+1.7)	64.9	48.1	27.0	47.0	55.3
Mask R-CNN w/HarmonicDet Loss ‡	ResNeXt-101-FPN	**46.9**	66.0	50.8	30.2	50.2	57.9

where N is the number of *Positive* samples. p_i and IoU_i denotes the predicted classification score and localization accuracy for the sample x_i, resp. Clearly, the smaller AIC value means the smaller prediction inconsistency, and vice versa. We visualize the AICs of the detectors trained with standard detection loss and our Harmonic loss in Fig. 4 (b). We can find that the AIC of the detector trained with Harmonic loss is significantly smaller than the model trained with the standard detection loss during the training. This also clearly means that our Harmonic loss can reconcile prediction inconsistency between classification and regression.
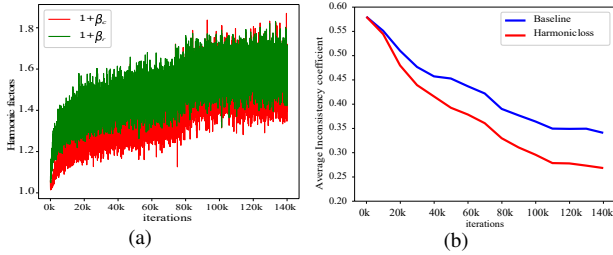

Figure 4. (a) is the distribution of the two harmonic factors (i.e., $1+\beta_r$, $1+\beta_c$) during the optimization process. (b) is the distribution of the AICs during the optimization process

4.5. Ablation Study

We perform a thorough ablation study on each component of our method. The Faster R-CNN detector is adopted to perform ablation experiments on MS COCO dataset. Table 3 shows the effects of each component of our method.

Table 3. Ablation results of each component (i.e., Harmonic loss and Harmonic IoU (HIoU) loss) of our methods on MS COCO `val` set. Faster R-CNN is adopted as the baseline.

method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline(Faster R-CNN)	37.3	58.6	40.7	21.0	40.1	49.2
+Harmonic loss	38.5	59.3	41.8	21.7	41.4	50.1
+Harmonic loss+IoU loss	38.7	59.3	42.1	21.6	41.8	50.3
+Harmonic loss+HIoU loss	**39.2**	59.2	42.7	22.0	42.0	50.9

Table 4. Ablation results of the TC loss in our Harmonic loss

Harmonic loss	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Without TC Loss	38.3	59.0	41.7	21.5	41.1	49.8
With TC Loss	38.5	59.3	41.8	21.7	41.4	50.1

Analysis of Harmonic Loss. As shown in Table 3, we first conduct the ablation experiments on the MS COCO dataset. Our Harmonic loss can improve the AP by 1.2% (from 37.3% to 38.5%). The detection result of each IoU level has been improved, and as the IoU increases, the improvement of AP is more significant. This indicates that the model trained with our Harmonic loss can output harmonious boxes with high co-occurrence of top classification and localization. We further analyze the effect of the TC loss in the Harmonic loss in Table 4. We can find that the APs can be improved by a small margin with low IoU thresholds, which indicates that the TC loss can convert some negative detection results with poor localization into positive detection results with IoU threshold greater than 0.5. Since the harmonic loss is proposed to reconcile classification task and regression task, but *does it really alleviate misalignment between the classification score and the localization accuracy in the testing phase*? In Fig. 5, we visualize the
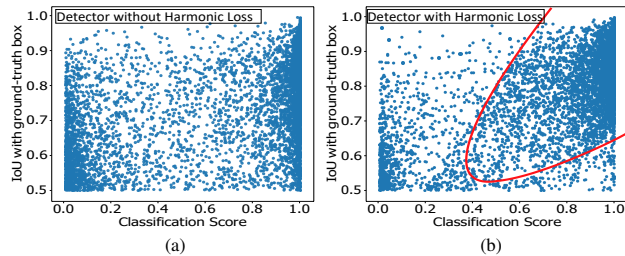
Figure 5. The distribution of classification scores and IoUs of all detected bboxes. The red curve is an indicator curve, and the detected bboxes at the upper right corner of the curve are harmonious detection results with high classification score and high IoU.

Table 5. Ablation results of the α and γ in HIoU loss.

α	γ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.5	0.8	38.9	59.3	42.1	21.9	41.7	50.5
1.0	0.8	39.0	59.3	42.3	22.1	41.7	50.8
1.5	0.8	**39.2**	59.2	42.7	22.0	42.0	50.9
2.5	0.8	38.9	58.6	43.0	21.6	41.8	50.6
1.5	0.0	38.7	59.3	42.1	21.6	41.8	50.3
1.5	0.5	38.9	59.2	42.4	21.7	41.8	50.7
1.5	0.8	**39.2**	59.2	42.7	22.0	42.0	50.9
1.5	1.0	39.0	58.7	42.8	21.9	41.9	50.9

distribution of classification scores and IoUs of all detected bounding boxes. Compared with detector without Harmonic loss (a), more predicted bounding boxes of our improved detector (b) appear inside the red curve, which means that our detector can predict more bounding boxes with both high classification scores and high regression accuracy (IoU). In other words, our detector clearly alleviates the disharmony between classification score and localization accuracy.

Analysis of HIoU Loss. In order to validate the effectiveness of our HIoU loss, we conduct three ablation experiments. *Firstly*, we add the standard IoU loss and HIoU loss to the Harmonic loss respectively. As shown in Table 3, when we add the standard IoU loss into Harmonic loss, the AP can only be improved by a small margin (from 38.5% to 38.7%). But when we add the HIoU loss into Harmonic loss, our method can achieve 0.7% improvement (from 38.5% to 39.2%) compared to the detector without HIoU loss. Specifically, the improvement for AP at higher IoU threshold (e.g., 0.75) is very significant, which means that our HIOU loss prevents the localization loss from being dominated by outliers in training phase. The reason is that our HIoU loss can alleviate the imbalance bias of samples with different IoU levels, as shown in Fig.3 (b). *Secondly*, we conduct experiments to analyze the effects of the focusing parameter γ of the HIoU loss. As shown in Table 5, we can get the best AP when we set the focusing parameter γ to 0.8. *Thirdly*, we also conduct experiments to analyze the effects of the trade-off α between SmoothL1 loss and HIoU loss. As shown in Table 5, we experiment with four different ratios ranging from 0.5 to 2.5 and get the best result when $\alpha = 1.5$. So we adopt $\alpha = 1.5$, $\gamma = 0.8$ in our experiments.
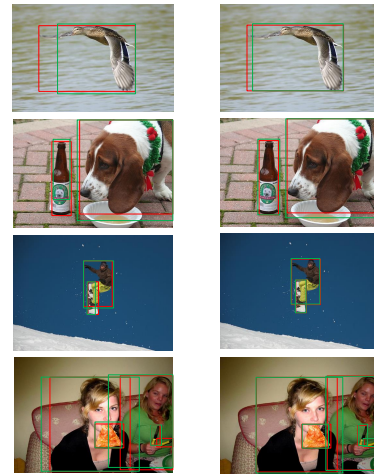

Figure 6. Qualitative comparisons between the standard detection loss and our HarmonicDet approach based on the Faster R-CNN. Several examples of irregular and occluded object detection are illustrated. The first column shows the results of Faster R-CNN trained with standard detection loss. The second column shows the results of Faster R-CNN trained with HarmonicDet loss. Red boxes represent the GT and green boxes represent the predicted boxes.

4.6. Qualitative Detection of Diverse Objects

Qualitative comparisons between the standard Faster R-CNN and our method are provided in Fig. 6. We can clearly find that, compared to Faster R-CNN trained with standard detection loss, the Faster R-CNN trained with our HarmonicDet loss can obtain more accurate detected bounding boxes. This indicates that by using HarmonicDet loss in the training phase to strengthen the correlation between the two tasks, the disharmony between classification score and localization accuracy can be effectively alleviated in the testing phase.

5. Conclusion

In this paper, in order to reconcile prediction consistency between classification and regression, we propose a Harmonic loss to harmonize the training of classification and regression in object detection. The Harmonic loss enables classification task and the localization task to dynamically promote each other in the training phase, thereby producing the consistent bounding boxes with high co-occurrence of top classification and localization in the testing phase. Furthermore, in order to prevent the localization loss from being dominated by outliers during training phase, a Harmonic IoU (HIoU) loss is proposed to harmonize the weight of the localization loss of different IoU-level samples. Comprehensive experiments by plugging and playing the proposed loss into six popular detectors demonstrate the generality and effectiveness of our HarmonicDet approach.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1, 7

[2] Yuhang Cao, Kai Chen, Chen Change Loy, and Dahua Lin. Prime sample attention in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[3] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018. 1

[4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 2

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 3

[6] Mark Everingham and John Winn. The pascal visual object classes challenge 2007 (voc2007) development kit. *University of Leeds, Tech. Rep*, 2007. 2

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1, 2

[8] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. 1

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[12] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 3

[13] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8577–8584, 2019. 3

[14] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6054–6063, 2019. 1

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1, 2

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 2, 5

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[18] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018. 2

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1, 2

[20] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2019. 3

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[24] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9627–9636, 2019. 3, 7

[25] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[26] Keyang Wang and Lei Zhang. Single-shot two-pronged detector with rectified iou loss. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1311–1319, 2020. 1

[27] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

[28] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 5

[29] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4203–4212, 2018. 1, 2

[30] Chengju Zhou, Meiqing Wu, and Siew-Kei Lam. Ssa-cnn: Semantic self-attention cnn for pedestrian detection. *arXiv preprint arXiv:1902.09080*, 2019. 1

[31] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 7

[32] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014. 2