Location-Aware Box Reasoning for Anchor-Based Single-Shot Object Detection

WENCHI MA, KAIDONG LI, AND GUANGHUI WANG, (Senior Member, IEEE)
Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS 66045, USA
Corresponding author: Guanghui Wang (ghwang@ku.edu)

ABSTRACT In the majority of object detection frameworks, the confidence of instance classification is used as the quality criterion of predicted bounding boxes, like the confidence-based ranking in non-maximum suppression (NMS). However, the quality of bounding boxes, indicating the spatial relations, is not only correlated with the classification scores. Compared with the region proposal network (RPN) based detectors, single-shot object detectors suffer the box quality as there is a lack of pre-selection of box proposals. In this paper, we aim at single-shot object detectors and propose a location-aware anchor-based reasoning (LAAR) for the bounding boxes. LAAR takes both the location and classification confidences into consideration for the quality evaluation of bounding boxes. We introduce a novel network block to learn the relative location between the anchors and the ground truths, denoted as a localization score, which acts as a location reference during the inference stage. The proposed localization score leads to an independent regression branch and calibrates the bounding box quality by scoring the predicted localization score so that the best-qualified bounding boxes can be picked up in NMS. Experiments on MS COCO and PASCAL VOC benchmarks demonstrate that the proposed location-aware framework enhances the performances of current anchor-based single-shot object detection frameworks and yields consistent and robust detection results.

INDEX TERMS Location-aware, box reasoning, calibrated quality score, single-shot, localization score, locscore head, object detection.

I. INTRODUCTION

Deep networks have been dramatically driving the progress of computer vision, bringing out a series of popular models for different vision tasks [31], [35], like image classification [3], [29], object detection [15], [32], crowd counting [25], depth estimation [10], and image translation [30]. Object detection plays an important role and serves as a prerequisite for numerous computer vision applications, such as instance segmentation, face recognition, autonomous driving, and video analysis [1], [9], [11], [12], [21]. In recent years, the performance of object detectors has been dramatically improved due to the advancement of deep network structure, well-annotated datasets, and effective optimization algorithms [17], [34].

In this paper, we aim at single-shot object detectors that yield a better trade-off between accuracy and speed,

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenbao Liu.

indicating a trend for future frameworks [19], [20]. We reveal the problem of an inadequate quality criterion for anchor-based bounding box candidates, which is very important for model optimization and detection evaluation. The reason lies in that the quality of bounding boxes should reflect both the spatial location accuracy and the classification probability. While as far as we know, in current deep learning-based object detection pipelines, the scores of the bounding boxes are shared with box-level classification confidence, which is predicted on the proposed features by the classifier. Most importantly, we cannot obtain the location assessment during the inference stage due to the deficiency of labels. It is insufficient to use the classification confidence to measure the bounding box quality since it only serves for distinguishing the semantic categories of proposals, while it is not aware of the assessment towards localization accuracy. The misalignment between classification confidence and bounding box quality is illustrated in Figure 1, from which we can see that, although the object instances obtain a high classification

FIGURE 1. Demonstrative detection results on MS COCO [18]. The predicted bounding boxes have high classification scores while the localization is misplaced or interceptive. The left two images show misplaced cases in which the zebra is located by a much larger region, and the car is not an actual object. The airplanes in the right two images are partially located, while all of them have high confidence scores. Our method predicts the spatial relation between box proposals and their possible targets so that the interception and misplacement can be minimized.

confidence score, the box-level localization is not unanimously accurate. If a predicted object is not scored properly, it might be mistaken as a false positive or negative, affecting the NMS process and leading to a decrease of average precision (AP). It is evident that the lack of effective scoring metrics towards the localization quality tends to impair the evaluation.

In this work, we focus on a more reasonable and effective scoring metric for anchor-based bounding box proposals. Different from most previous works that either pursue high-quality classification boxes or focus on score correction working on two-stage object detectors, we demonstrate that there is room of further improvement for popular anchor-based single-shot object detection models by introducing calibrated quality scores that take the location confidence into consideration. Compared with RPN-based frameworks, single-shot object detectors highly depend on qualified box proposals as there is no pre-sift scheme. As anchor-based methods, they are sensitive to location information, which brings challenges for box sifting. To solve this issue, we propose a calibrated quality score (**CQS**) for each box proposal to realize the location awareness. The localization score indicates the spatial relation to its most-probable target ground truth and ranks the proposals based on the calibrated quality score rather than the classification confidence.

In anchor-based single-shot object detection, the bounding box proposals are regressed by the space shift relative to the anchors. Thus, the spatial relationship of an anchor and an object ground truth depicts an expectation or estimation of the location relationship between the corresponding box proposal and the target, as depicted in Figure 2. Inspired by the Average Precision (AP) metric of object detection using pixel-level Intersection-over-Union (IoU) between the predicted bounding box and its ground truth to describe the

quality of predictions, we propose a network module to learn the IoU between the anchors and the ground truth directly. For the convenience of discussion, we call it **AIoU**. We adopt the proposed locscore to learn this AIoU during the training time, and when given the locscore in the test phase, the quality of bounding boxes is reevaluated by integrating locscore into the classification confidence so that the reasoned box proposals are aware of both the location information and the semantic categories.

Compared with localization and classification regressions that take the ground truths from the labeled dataset, the learning for AIoU only needs to calculate the IoU between the anchors and the ground truths as a target, without further labeling the dataset. Within a detection model, we implement the locscore prediction network as the locscore head, which takes the feature outputs and the calculated AIoU as inputs, and is trained with a common regression loss. We implement object detection experiments with the proposed location-aware anchor-based box reasoning module on popular single-shot object detectors. The results demonstrate that our method can promote the performance of object detection and yield consistent and robust detection results. The main contributions of our paper include:

C1. We propose a novel bounding box reasoning method that is aware of the spatial relationship between the box proposals and the probable target ground truth. It is one of the first algorithms that address the issue caused by scoring bounding box proposals only by the classification probability.

C2. This is the first location-aware detection framework designed for the single-shot networks that naturally take the pools of anchor-based box proposals as candidates, ensuring a one-shot learning fashion.

C3. The proposed plug-in locscore head can be integrated with any single-shot detection networks and regressed easily in an end-to-end fashion. By calibrating the detection quality with locscore, the bounding boxes can be penalized if it has high classification confidence while relatively poor localization accuracy.

C4. We demonstrate the effectiveness of the location-aware anchor-based box reasoning scheme through extensive experiments. By introducing the proposed calibrated quality score into the evaluation metric of box proposals, the detection performance is further improved.

II. RELATED WORK
A. OBJECT DETECTION
Multiclass object detection is a core task in the context of deep learning based computer vision projects, which is the joint work of the classification towards contents and the localization towards bounding boxes of instances. Most of these methods adopt the CNN [14] based bounding box and classification regressions, followed by a Non-Max Suppression (NMS) algorithm to sift best-qualified box proposals.

Bounding box regression was first introduced in R-NN [8]. It enables regions of interest (ROI) to estimate the updated
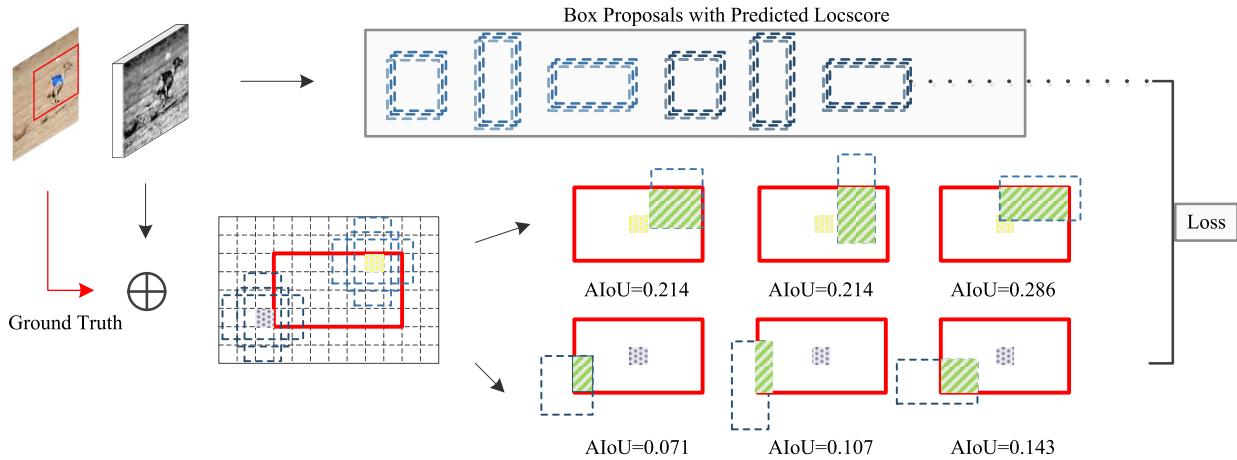
FIGURE 2. AIoU Definition and the Locscore illustration. AIoU as the target of the proposed locscore only needs the input image and its corresponding ground truth. In the fashion of convolution, we evaluate the default boxes with multiple scales at different anchor bases, as the yellow and purple color lumps shown in this figure. For each default box (represented by the blue dotted rectangular), in addition to predicting the shape offsets and the confidence scores as conventional detectors, we also predict the locscore which assesses the possibility of how close the object is to the ground truth. The locscore is learned towards the AIoU calculated by the anchor box and the ground truth, which is denoted by the red angular box in this figure. Specifically, the locscore of a box proposal is learned to match the AIoU between its corresponding anchor box and a certain ground truth box.

bounding boxes with the purpose of better matching the nearest object instance. Prior works, from Fast R-CNN, Faster R-CNN [24], R-FCN [5], to YOLO [23], SSD [19], RetinaNet [17], and RefineDet [33], have demonstrated that the detection task can be improved with multiple bounding box regression stages [7], flexible anchor matching [34], the increase of the number of anchors, and the enlargement of the input image resolution, including image pyramids [16]. Among them, the most widely-used and efficient technique is the anchor-based multibox algorithm that can handle scale variation, one of the challenging problems for one-shot object detection. Anchor boxes are designed for discretizing the continuous space of all possible instance boxes into a finite number of boxes with predefined locations, scales, and aspect ratios [36]. The created instance boxes are regressed to match the ground truth bounding boxes based on the Intersection-over-Union (IoU) overlap, by location shift at certain base anchor with the predefined locations.

However, there exist underlying limitations. On one hand, the quality of the proposed bounding boxes is only measured by the classification score, leading to the misalignment between the box score and box quality. Due to the unreliability of the box score, a proposal with higher IoU against the ground truth will be ranked with low priority if it obtains lower classification confidence. In this situation, the Average Precision (AP) can be degraded. On the other hand, compared with RPN [24], the anchor-based technique is more sensitive to box quality especially for the consistency of classification confidence and location accuracy since there is no pre-sift scheme for box proposals in the one-stage case.

B. DETECTION SCORING AND CORRECTION
The misalignment of the box score and actual quality has aroused much attention and several correction methods have

been proposed in recent years. Tychsen-smith and Petersson [27] presented a Fitness NMS that corrects the detection score by learning the statistics of best matching detected bounding boxes with the ground truth as a corrective factor. It formulates box IoU statistics prediction as the classification task. It is specifically designed for Denet [26], which restrains its application to arbitrary object detection frameworks.

Jiang *et al*. [13] proposed a standalone IoU-Net which is based on a similar R-CNN structure with a proposal pre-sift scheme to predict IoU between the predicted boxes and the ground truths. It manually designs bounding box filtering as an addition to the data pool of box proposals. The IoU-guided NMS ranks bounding boxes by the predicted localization confidence rather than the conventional classification confidence. Cheng *et al*. adopted a separate network to correct the scores of samples by processing false-positive samples [4]. SoftMax [2] proposed to use the overlap between two boxes to correct the low score box. Neumann *et al*. [22] proposed a relaxed softmax to predict the temperature scaling factor in standard softmax for safety-critical pedestrian detection. Both of the two approaches are designed for the two-stage R-FCN based models, relying on the clean proposal data pool. Wu *et al*. proposed the IoU-aware approach scores the location results while it is merely a RetinaNet based detector [28], not an arbitrary method.

Different from the above methods, this study focuses on the essence of the evaluation towards anchor-based box reasoning in single-shot frameworks. We assign each predicted box with a location score by taking aware of the spatial relation between its based anchor and the ground truth. The proposed approach takes both the classification confidence and the location accuracy into consideration to create a complete evaluation of instance box quality so as to narrow the gap between the box score and the actual quality. Furthermore,

we build an independent regression branch in the single-shot object detection framework that learns the location confidence specifically and merges this information into the box quality evaluation metric of NMS so as to obtain a more reliable priority ranking.

III. LOCATION-AWARE BOX REASONING

A. MOTIVATION

In current object detection frameworks, the classification and localization regressions are taken as two independent processes. The evaluation towards a detection hypothesis, the detected bounding box, is determined by the highest-ranked element in the classification scores. However, there exist certain situations where the predicted box with a high classification score has low localization accuracy. This kind of hypothesis is harmful in most detection evaluation protocols, such as MS COCO. It is important that a detector can determine when the detection results are trust-worthy and when they are not. This motivates us to integrate the localization score by location-aware anchor-based reasoning for every predicted bounding box based on an anchor position.

Most previous methods do not consider location confidence as one of the evaluation factors that contribute to the box quality [4], [27] [2], [22] and the majority of them are designed as specific detectors or merely applied for two-stage detectors with the pre-sift scheme. Although IoUNet obtains competing results with the proposed IoU-guide NMS algorithm that takes localization confidence into consideration, it ranks the boxes only by the localization confidence which highly relies on the clean data pool of proposals produced by two-stage detection models. It is hard to apply it in a single-shot fashion. We propose the location-aware anchor-based box reasoning that focuses on arbitrary single-shot detectors for a better trade-off between accuracy and efficiency.

We instantiate the location-aware reasoning module by showing how to apply it to the anchor-based single-shot detectors. Without loss of generality, we apply the proposed module to the state-of-the-art RetinaNet and SSD with an additional Locscore head that learns the IoU between the anchor and ground truth, and demonstrates our design from the following aspects: 1) how to realize location awareness for anchors; 2) how to create the branch of location score in the network; and 3) how to generate location-aware anchor-based reasoning during inference time.

B. LOCATION AWARENESS

From the perspective of conception, location awareness is simple. In anchor-based detectors, the introduction of location awareness supplements the evaluation towards the quality of the bounding boxes from the perspective of location accuracy. It is realized by learning the IoU between the anchors and ground truths, producing the localization scores.

Localization Score: We begin with briefly reviewing the evaluation metric towards the bounding box proposals. Following the anchor-based detection, the proposals are created based on anchors with various scales at different positions on the feature maps. The network extracts features from the backbone and performs proposal classification and bounding box regressions respectively. The former one yields confidence scores regardless of the location reference, while the latter one regresses the space migration of candidates. Although the predicted bounding boxes with classification confidence and location prediction, the quality of the box candidates can only be evaluated by the confidence score, without localization assessment. This is due to the lack of ground truth as a reference during the inference stage. Thus, there is a gap between the current metrics and the actual need for evaluating the quality of box proposals.

We define $P(lc|a_j)$ as the localization score by learning the pixel-level IoU between bounding box a_j and any object ground truth b_j. We define it as "AIoU".

$$AIoU = IoU(a_j, b_j) \qquad (1)$$

In ideal anchor-based detectors, the object is detected by three elements: anchor, box proposal, and object ground truth. Conventionally, we build the direct correspondence between the anchors and the boxes by regressing space shift, and the relations between the boxes and the object ground truths by matching features. However, there is no direct depiction of the relationship between the anchors and object ground truths, where an underlying location link exists. The introduction of the location score complements the relational structure of the three essential elements and further calibrates the quality criterion of bounding box proposals by the definition expressed as below.

$$S(a_j) = P(c|a_j) \cdot P(lc|a_j) \qquad (2)$$

where $P(c|a_j)$ denotes the confidence score, and $S(a_j)$ is defined as the calibrated quality criterion of the bounding box proposals. Thus, $S(a_j)$ should work well on two tasks: indicating the right category that the box belongs to and regressing the IoU of the proposals and the foreground objects.

C. LOCALIZATION SCORE REGRESSION

Locscore Head: Conventionally, classification and regression are two independent branches for all object categories. Without loss of generality, we introduce the Locscore Head as the third independent branch to predict the IoU between anchors and object ground truths. This head simply follows the layer structure of the other two existing heads so as to save the network characteristics and the advantages of the framework. Thus, it can be implemented as a plug-in module and be integrated with any arbitrary single-shot object detection model.

The Locscore Head receives the concatenation of features from the output layer of the network as its inputs. It predicts the localization score for each anchor box on the feature maps, which depicts the location relations between the box and the target ground truth. In anchor-based object detection, each predicted bounding box is created based on a certain anchor and regressed location migration. Thus, according to
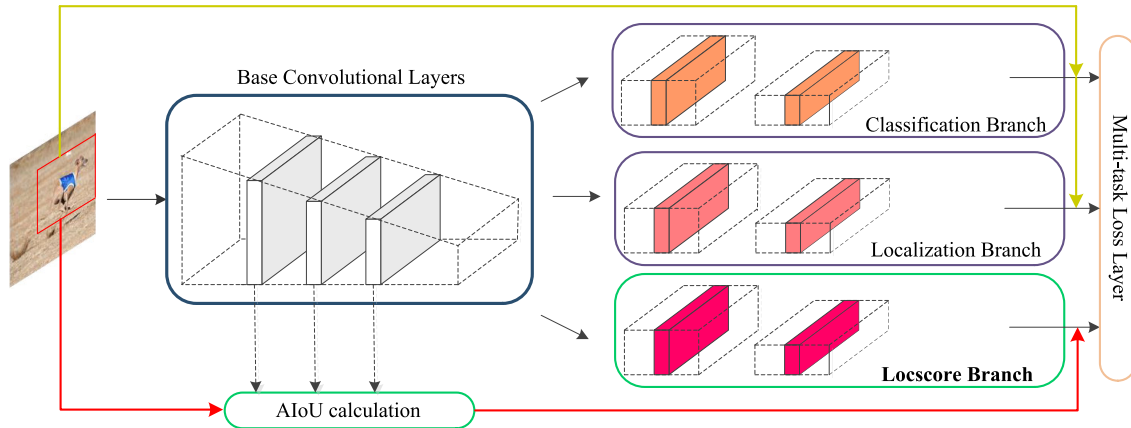
FIGURE 3. The network architecture of object detection with location-aware anchor-based reasoning. The input image is fed into the backbone network to generate feature maps with RoI information. The Locscore Branch is the standard component of the improved model. It takes the output features from the backbone network as inputs and provides a predictive locscore at the end, where its layer structure just follows the one of the classification branch or the localization branch.

this correspondence among the three elements, each predicted bounding box proposal is corresponding to one anchor so that the box would be assigned a localization score indicating the maximum possibility that it is related to an object from the perspective of localization.

Based on the analysis above, since the Locscore Head shares the same concatenated features with the other two Heads, classification, and box regression, the predicted three elements have an inner congruent relationship. The Locscore Head, thus, can be taken as an independent regression branch and treated as an individual learning task.

We define the Locscore Loss to regress the Localization Score. It follows the loss definition for classification regression. Then the Locscore Head is integrated into an anchor-based object detection framework, and the whole network can be trained end-to-end. Specifically, we define the classification loss and box regression loss as L_{cl} and L_{bb}, respectively. In addition, we introduce locscore loss L_{lc} as another penalty item to the cost function, as shown below,

$$L^* = \lambda_1 * L_{cl} + \lambda_2 * L_{bb} + \lambda_3 * L_{lc} \qquad (3)$$

where L^* denotes the final loss function. In all experiments, we adopt equal weights for the three loss items in consideration for stability, so that $\lambda_1 = \lambda_2 = \lambda_3 = 1$. The locscore loss forms a lower bound in the space localization, and by training, we further pull down this lower bound.

D. BOX REASONING DURING INFERENCE
1) CALIBRATED QUALITY SCORE (CQS)
We define and propose a calibrated quality score by introducing the localization confidence into the assessment of the predicted bounding box proposals. Thus, the quality score is disintegrated into two data spaces, where both classification confidence and location accuracy are taken into consideration, as shown in Equation 2. This CQS during inference time becomes a new criterion for the sifting of qualified

candidates. It complements the defect of independent classification and location regressions that lead to the deficiency of localization reference during inference time.

2) INFERENCE
At the inference stage, the proposed candidates with coordinates, conferences, and localization scores are integrated into the non-maximum suppression (NMS) algorithm. Different from its classical counterparts, the NMS in our model does not rank the box proposals merely by classification probabilities in the first step. Instead, we use the CQS as the ranking criterion so as to push the box candidates to indicate the spatial relations with potential objects, in terms of the initial idea that the quality of bounding boxes is tightly correlated to both the spatial information and classification confidence. We assume the saved boxes after the above sifting are qualified candidates. But in order to weaken the sensitivity towards the less qualified proposals for the single-shot models, we adopt the confidence cluster [13] to further enhance the reliability of the sifted boxes by updating the confidence score S_i of box i with $S_i = max(S_i, S_j)$, where j indicates box j that is deleted by box i in NMS. Details are shown in Algorithm 1.

Specifically, suppose the network outputs N bounding boxes, the NMS firstly rank them by the proposed CQS, then we follow the same procedure to remove the candidate boxes which overlap each other over a threshold of $\epsilon = 0.5$. At last, the top-k scored boxes are selected and fed into the output head to generate multi-class boxes.

IV. EXPERIMENTS
We conduct experiments on the detection tasks of the MS COCO [18] and PASCAL VOC [6] datasets. MS COCO contains 80 object categories, we follow COCO 2017 settings, using the 115k images *train* split for training, 5k *validation* split for results analysis. The COCO results are reported by

---

Algorithm 1 Location-Aware Box Reasoning. Classification Confidence and Localization Score Are Independently Regressed During the Training Time but the Two Values Are Taken as Combined Consideration During the Inference Time When Evaluating the Anchor-Based Box Proposals

---

Input: B, P_c, P_{lc}, ϵ.
B: set of anchor-based bounding box proposals.
P_c: classification confidence by mapping f_c.
P_{lc}: localization score by mapping f_{lc}.
ϵ: IoU threshold in NMS.
Output: D, set of detected boxes with classification confidence P_c.
1:    $D \leftarrow \varnothing$
2:    **while** $B \neq \varnothing$ **do**
3:        $S \leftarrow P_c \times P_{lc}$
4:        $b_m \leftarrow \arg \max_i S(b_i)$
5:        $S_m \leftarrow S(b_i)$
6:        $P_m \leftarrow P_c(b_i)$
7:        $B \leftarrow B \setminus b_m$
8:        **for** $b_j \in B$ **do**
9:           **if** $\text{IoU}(b_m, b_j) > \epsilon$ **then**
10:             $B \leftarrow B \setminus b_j$
11:             **if** classification cluster **then**
12:                $P_m \leftarrow max(P_c(b_j), P_m)$.
13:        $D \leftarrow D \cup \{\langle b_m, P_m \rangle\}$

---

its evaluation metrics AP (Average Precision over IoU thresholds), including, AP@0.5 (IoU equals to 0.5), AP@0.75 (IoU equals to 0.75), AP (averaged on AP over IoU thresholds from 0.5-0.95 with a step size of 0.05), AP_S, AP_M, AP_L (AP at different scales of objects). In the VOC experiments, we follow the same practice as in the literature, the models are trained on the union of PASCAL VOC2007 and 2012 trainval set (16,551 images) and tested on PASCAL VOC 2007 test set (4952 images). The overlap threshold for each one of the 20 categories in VOC is set to 0.5.

A. IMPLEMENTATION DETAILS

We adopt the consistent location-aware box reasoning (LAAR) framework for all experiments. We use the ResNet-50 based FPN network and VGG-16 based SSD network for ablation study, respectively. For ResNet-50 FPN, the input images are resized with a minimum 608px along the short axis and a maximum of 1024px along the long axis for both training and test. We train the network and choose the model at the epoch that yields the best performance. The learning rate is reduced by the Plateau strategy, the same as that of the original RetinaNet. For SSD, the input images are resized to 300 × 300 for both training and test, as the common rule in literature. The rest of all configurations are identical to the realization in [19]. We train the network for 120,000 iterations and decrease the learning rate after 80,000 and 100,000 iterations. The optimizer for RetinaNet experiments is Adam with an initial learning rate of 0.00001,

and for SSD experiments is SGD with momentum 0.9. In the test, all the results are evaluated by the NMS, where the top-100 score detection is retained for each image.

Learning Scenarios: In order to identify the gains of locscore regression constraint and location-aware box reasoning respectively, we intentionally design independent learning scenarios where we do solo locscore regression constraint without quality score calibration, the complete LAAR detection framework with CQS, and the complete LAAR detection with CQS and confidence cluster. We list them as follows:

- *Independent Locscore Constraint (ILC):* We introduce the locscore regression during training time while do not consider the predicted locscore to calibrate the classification score during the inference period.
- *Locscore Constraint with CQS (LC):* We conduct complete location-aware box reasoning with the given detector, which means we add locscore regression constraint during training and introduce the calibrated quality score by the predicted locscore during the test time.
- *Locscore Constraint with CQS and classification cluster (LC+CS):* Based on LC, we introduce the classification cluster after calibrating quality score by the predicted locscore, as shown in Algorithm 1.

B. ABLATION STUDIES

We evaluate the contribution of one important element to our location-aware box reasoning for object detection, the constraint brought by the Localization Score Regression.

Locscore Constraint for Better Optimization: Compared with conventional object detection frameworks, we introduce the additional constraint term in the loss function by doing the localization score regression. As far as we know, this is the first time to directly explore the relationship between the anchors and the ground truths in single-shot fashions. The proposed procedure explores the predefined prior information of anchor boxes and the ground truths from the perspective of spatial location. As is known, anchor-based fashion defines fixed anchor positions and their multiple-ratio variations on a feature map, which means for a certain image, there exist predefined location relations between the anchor boxes and the ground truths. We introduce this relation in the penalty function to help constrain the classification and localization regressions. Especially for the latter one during training, when the coordinates are learned in a deflected direction, there exists a correction by the locscore regression. This constraint results in better optimization as demonstrated in Table 1 and 2.

In Table 1, the ILC version leads the original RetinaNet in most cases. For $\text{AP}_{0.75}$, ILC improves the detection accuracy by a rough 1% than the original ResNet. Although the ILC falls behind AP_S, it leads by large margins in both AP_M and AP_L. In Table 2, we obtain consistent results. In this VOC setting, we list the AP results for all the 20 categories. The ILC shows better mAP than the original SSD and yields clearly higher AP for most categories. These results support the idea that the introduction of locscore regression yields effectively

TABLE 1. The mAP of RetinaNet on COCO val2017. R-50 indicates ResNet50 with FPN and R-Net refers to RetinaNet.

	R-Net	R-Net+ILC	R-Net+LC	R-Net+LC+CS
Backbone	R-50	R-50	R-50	R-50
AP	30.4 %	30.9 %	30.7 %	**30.9** %
$AP_{0.5}$	47.3 %	**47.7** %	47.2%	47.4 %
$AP_{0.75}$	32.1 %	33.0 %	32.9%	**33.2** %
AP_S	**13.9**%	13.0 %	12.9 %	13.0 %
AP_M	33.1 %	**34.0** %	33.8%	33.9 %
AP_L	43.7%	44.1 %	44.2%	**44.3** %

TABLE 2. mAP of SSD on VOC2017. The category name indicates its corresponding AP result.

	SSD	SSD+ILC	SSD+LC	SSD+LC+CS
Backbone	VGG-16	VGG-16	VGG-16	VGG-16
aeroplane	82.97 %	81.87 %	82.43 %	82.41 %
bicycle	84.18%	83.61 %	83.13%	83.18 %
bird	75.34 %	75.23 %	74.23 %	75.56 %
boat	70.98%	70.35 %	70.62 %	70.71 %
bottle	50.44 %	51.97 %	51.62%	51.90 %
bus	84.29%	86.05 %	84.49%	86.03 %
car	86.32%	85.30 %	84.60%	85.53 %
cat	88.12%	88.36 %	87.45%	88.26 %
chair	61.54%	62.18 %	58.82%	62.02 %
cow	79.94%	83.03 %	82.28%	83.18 %
diningtable	77.12%	75.80 %	72.40%	76.20 %
dog	85.05%	84.31 %	82.38%	84.13 %
horse	87.60%	87.03 %	85.92%	87.68 %
motorbike	82.84%	82.95 %	81.83%	83.50 %
person	78.98%	79.09 %	77.71%	79.12 %
pottedplant	52.17%	51.94 %	50.27%	51.45 %
sheep	78.61%	77.40 %	75.40%	77.05 %
sofa	78.33%	80.29 %	77.49%	80.12%
train	87.88%	86.65 %	84.80%	87.71%
tvmonitor	76.33%	77.29 %	74.93%	77.10%
mAP	77.45%	77.53 %	76.14%	**77.64%**

positive constraint towards the other two regressions. Thus, the locscore branch boosts the mutual promotion of classification and localization, helping to improve the situation when confidence score and localization accuracy are opposite, such as when confidence score is 0.6 and localization accuracy is 0.2 compared to that when confidence score is 0.2 and localization accuracy is 0.6, which guarantees the feasibility of the proposed approach.

C. QUANTITATIVE RESULTS

We extensively evaluated the proposed method with two popular detectors, SSD and RetinaNet, on Pascal VOC 2007 and COCO val2017, respectively. We also compared the proposed algorithm with state-of-the-art methods, SoftMax [27] and IoUNet [13]. Since these methods are confined to the application of R-CNN based two-stage frameworks, they are highly dependent on the pre-sift scheme by RPN, where the proposal pool could already be regarded as clean data. The proposed method aims at anchor-based single-shot detection models without the pre-sift scheme, where the anchor-based produced boxes can be regarded as rough candidates. We conduct experiments and verify that we can not directly introduce the above algorithms in the single-shot fashion as a comparison, like the one proposed by IoUNet by ranking the boxes using localization confidence. It leads to a sharp

drop in mAP. It is understandable that the produced boxes by anchors are rough candidates that are unreliable for quality ranking. Therefore, these methods can not be directly applied to single-shot models.

To make a fair comparison, we integrate the core idea of IoU-guided NMS produced by IoUNet to merge the classification cluster into our algorithm and form the 'LC+CS', described in lines 11 and 12 of Algorithm 1. We then conduct comparison and show quantitative results in Table 1 and 2. From Table 1 we can see that, compared with RetinaNet, LC model achieves stable improvement in most cases. Specifically, LC obtains better AP, $AP_{0.75}$, AP_M, and AP_L. For $AP_{0.75}$, LC achieves an enhancement of 0.8%, and for AP_M, LC improves by a margin of 0.7%. We can conclude that, for the detection accuracy with high request and objects with the most common sizes, our method exhibits clear advantages. In Table 2, the LC version does not lead the ranking although its corresponding ILC version performs better than the original SSD model. This reflects the fact that the anchor-based single-shot models are sensitive to location accuracy and the produced rough candidates have less reliable location outputs than the expectation. From Tables 1 and 2 we can see that 'LC+CS' achieves the best results in both experiments, with over 1% enhancement in some cases, such as in RetinaNet at $AP_{0.75}$. The results demonstrate that the introduction of classification clustering could compensate for the uncertainty caused by the location outputs.

D. COMPARISON AND DISCUSSION

In this section, we first discuss the quality of the predicted bounding boxes and investigate the upper bound of the performance of the LAAR model, and analyze the benefits of locscore learning. Here, all the results are evaluated on COCO2017 validation set using RetinaNet and ReinaNet with the LC models.

1) FITTER AND TIGHTER BOUNDING BOXES

In Figure 4, It is evident that the LC model predicts higher-qualified boxes than RetinaNet, and most of the boxes are tighter. Specifically, the boxes for the vase, the eagle, and the truck, are more accurate, and the box for the chair shows better performance in occluded cases. This demonstrates that the introduction of locscore learning can improve the accuracy of bounding boxes, and help select the best proposals that have the maximum alignment between the quality score and box quality. Tighter boxes in practice can help clear up some current dilemmas in industry.

2) BETTER FOR HARD OBJECTS

In Figure 5, it can be seen that the LC model is able to detect harder objects, like the small backpacks, the occluded cellphone, and the persons in the audience. These small or occluded objects are easy to be overlooked during the suppression process, as their classification scores can be small due to the deficiency of the effective features. While with the locscore, the detection score can be calibrated as these
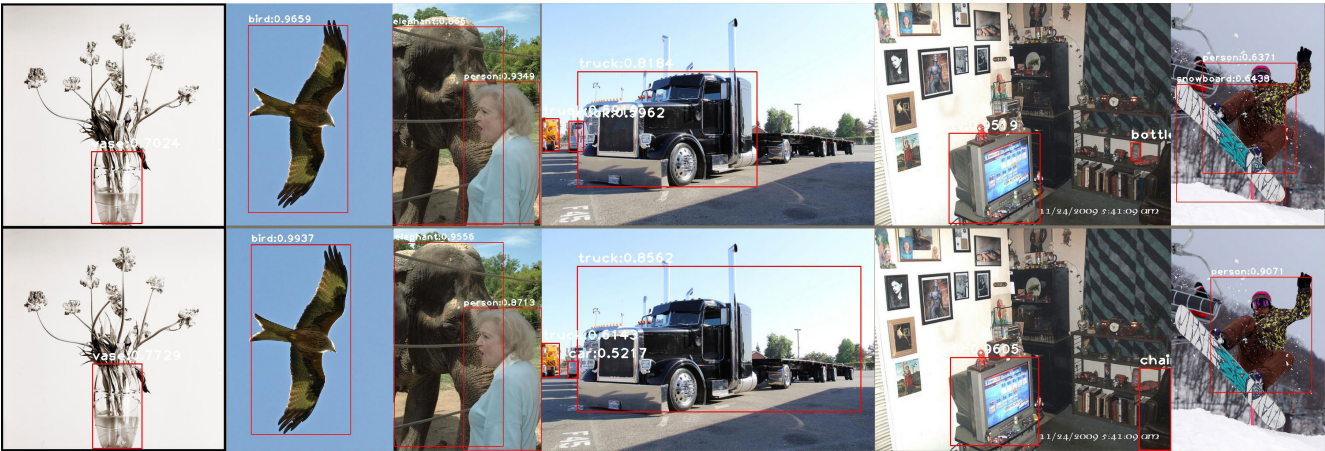
FIGURE 4. COCO cases compared between RetinaNet (1st row) and RetineNet+LC (2nd row), where LC version achieves higher localization accuracy.

FIGURE 5. COCO cases compared between RetinaNet (1st row) and RetineNet+LC (2nd row), where LC version performs better for hard objects.
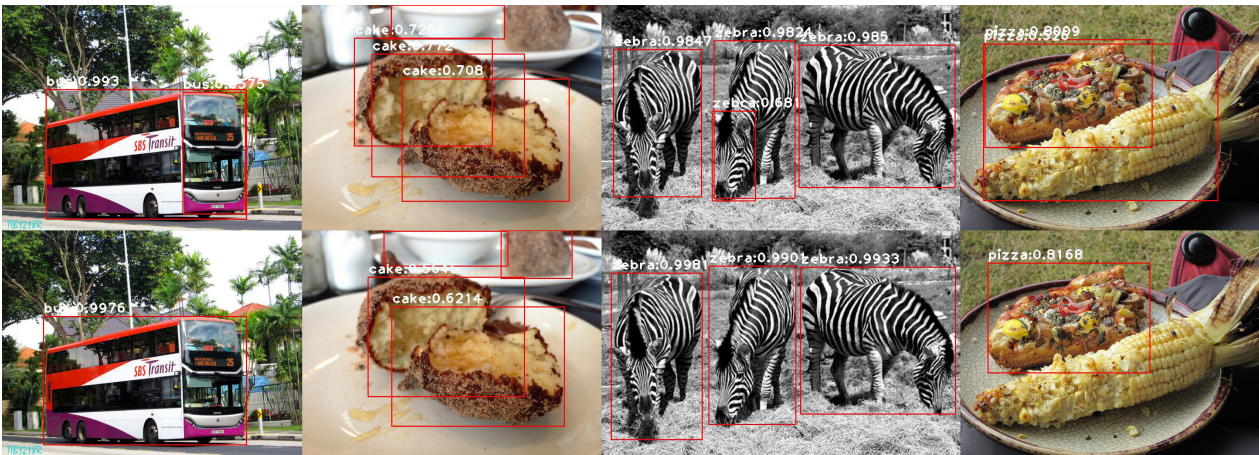

FIGURE 6. COCO cases compared between RetinaNet (1st row) and RetineNet+LC (2nd row), where LC version has better filtering for low-quality boxes.

hard objects could probably have high localization scores if they are labeled. Thus, the actual detection accuracy can be raised by calibrating the box quality score especially when the original classification confidence is low.

3) WAIVING LOW-QUALITY BOXES

In addition, from Figure 6, we can see the LC model is better at sifting out some low-qualified boxes, like the smaller bus box, the redundant cake, and zebra box. Similarly, this

can be explained by the calibrated quality score. Although some boxes' classification scores are high, if their location scores are low, these boxes can still be regarded as low-quality objects, which could be disregarded in NMS.

V. CONCLUSION

This paper reveals the problem of object detection score as one of the primary limitations of current anchor-based single-shot object detectors. To address this issue, we have proposed the localization score (locscore) regression and location-aware box reasoning, where the classification score is aligned with the predicted locsocre so that the localization accuracy is taken into the assessment of the quality of the bounding box proposals, which has been overlooked in most popular object detection frameworks. Extensive experimental results show that the proposed approach can consistently improve the detector's performance to yield reliable bounding boxes. The proposed module can be directly applied to any single-shot object detection models to improve their performance in both classification and localization.

REFERENCES

[1] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides, "Faster than real-time facial alignment: A 3D spatial transformer network approach in unconstrained poses," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3980–3989.

[2] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5561–5569.

[3] F. Cen and G. Wang, "Boosting occluded image classification via subspace decomposition-based estimation of deep features," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3409–3422, Jul. 2020.

[4] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting RCNN: On awakening the classification power of faster RCNN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 453–468.

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[7] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1134–1142.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[10] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4676–4689, Sep. 2018.

[11] Y. He, X. Ma, X. Luo, J. Li, M. Zhao, B. An, and X. Guan, "Vehicle traffic driven camera placement for better metropolis security surveillance," 2017, *arXiv:1705.08508*. [Online]. Available: http://arxiv.org/abs/1705.08508

[12] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6409–6418.

[13] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 784–799.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[15] K. Li, W. Ma, U. Sajid, Y. Wu, and G. Wang, "Object detection with convolutional neural networks," 2019, *arXiv:1912.01844*. [Online]. Available: http://arxiv.org/abs/1912.01844

[16] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[18] 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 *Proc. Eur. Conf. Comput. Vis.* Seoul, South Korea: Springer, 2014, pp. 740–755.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2016, pp. 21–37.

[20] W. Ma, Y. Wu, F. Cen, and G. Wang, "MDFN: Multi-scale deep feature learning network for object detection," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107149.

[21] W. Ma, Y. Wu, Z. Wang, and G. Wang, "MDCN: Multi-scale, deep inception convolutional neural networks for efficient object detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2510–2515.

[22] L. Neumann, A. Zisserman, and A. Vedaldi, "Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection," in *Proc. NIPS*, 2018, pp. 1–8.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[25] U. Sajid, H. Sajid, H. Wang, and G. Wang, "ZoomCount: A zooming mechanism for crowd counting in static images," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Mar. 5, 2020, doi: 10.1109/TCSVT.2020.2978717.

[26] L. Tychsen-Smith and L. Petersson, "DeNet: Scalable real-time object detection with directed sparse sampling," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 428–436.

[27] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness NMS and bounded IoU loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6877–6885.

[28] S. Wu, X. Li, and X. Wang, "IoU-aware single-stage object detector for accurate localization," *Image Vis. Comput.*, vol. 97, May 2020, Art. no. 103911.

[29] Y. Wu, Z. Zhang, and G. Wang, "Unsupervised deep feature transfer for low resolution image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–5.

[30] W. Xu, S. Keshmiri, and G. Wang, "Adversarially approximated autoencoder for image generation and manipulation," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2387–2396, Sep. 2019.

[31] W. Xu, G. Wang, A. Sullivan, and Z. Zhang, "Towards learning affine-invariant representations via data-efficient CNNs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 904–913.

[32] W. Xu, Y. Wu, W. Ma, and G. Wang, "Adaptively denoising proposal collection for weakly supervised object localization," *Neural Process. Lett.*, vol. 51, no. 1, pp. 993–1006, Feb. 2020.

[33] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.

[34] X. Zhang, F. Wan, C. Liu, R. Ji, and Q. Ye, "Freeanchor: Learning to match anchors for visual object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 147–155.

[35] Z. Zhang, W. Ma, Y. Wu, and G. Wang, "Self-orthogonality module: A network architecture plug-in for learning orthogonal filters," 2020, *arXiv:2001.01275*. [Online]. Available: http://arxiv.org/abs/2001.01275

[36] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," 2019, *arXiv:1903.00621*. [Online]. Available: http://arxiv.org/abs/1903.00621
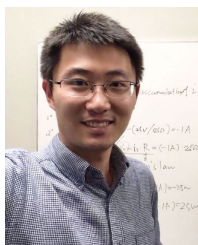
WENCHI MA received the master's degree from the Harbin Institute of Technology, Harbin, China, in 2015. He is currently pursuing the Ph.D. degree with The University of Kansas, Lawrence, KS, USA. His research interests include deep learning, computer vision, pattern recognition, and image processing.

KAIDONG LI received the bachelor's degree from Shanghai Jiao Tong University, China, and the master's degree from the University of Southern California, CA, USA. He is currently pursuing the Ph.D. degree in computer science with The University of Kansas, with a focus on computer vision and machine learning

GUANGHUI WANG (Senior Member, IEEE) is currently an Associate Professor with The University of Kansas, USA. He has authored one book *Guide to Three Dimensional Structure and Motion Factorization* (Springer-Verlag). He has published over 120 papers in peer-reviewed journals and conference proceedings. His research interests include computer vision, structure from motion, object detection and tracking, artificial intelligence, and robot localization and navigation. He has served as an Associate Editor and on the Editorial Board of two journals, as an Area Chair or a TPC Member of over 30 conferences, and as a Reviewer for over 20 journals.

● ● ●