Noise-Aware Evaluation of Object Detectors

Jeffri Murrugarra-Llerena
Computer Science Department
Stony Brook University
jmurrugarral@cs.stonybrook.edu

Claudio R. Jung
Institute of Informatics
Federal University of Rio Grande do Sul
crjung@inf.ufrgs.br

Abstract

Supervised object detection requires annotated datasets for training and evaluation purposes. However, human annotation of large datasets is error-prone, and frequent mistakes are erroneous labels, missing objects, and imprecise bounding boxes. The main goals of this work are to quantify the extent of annotation noise in terms of corner-wise discrepancies, assess how it impacts evaluation metrics for object detection, and propose noise-aware alternatives that serve as upper and lower bounds for a baseline metric. We focus our analysis on the Microsoft COCO dataset and re-evaluate several state-of-the-art object detectors using the proposed metrics. We show that the Average Precision (AP) metric might be considerably over or under-estimated, particularly for small objects and restrictive IoU acceptance thresholds. Our code is available at `https://github.com/Artcs1/Error-Aware`.

1. Introduction

Object detection is a core component in many high-level tasks, such as autonomous driving, surveillance, healthcare, and object tracking. With the widespread adoption of deep learning for supervised object detection [34], several datasets have been proposed in the past years, such as VOC [5], COCO [13], LVIS [7], Open Images Dataset (OID) [11], DOTAv1.0-2.0 [4], to name a few.

The size of available datasets has also increased. For example, OID v4 [11] presents over 1.7M images and 16M bounding boxes in the train split. At this scale, the evaluation of object detectors has relied more and more on *objective* quality metrics, and the most popular is the mean average precision (AP). This metric explores the Intersection-over-Union (IoU) to assess if the predictions made by the detector match a ground truth (GT) annotation based on an acceptance threshold, but the choice for the IoU threshold or even the use of the IoU itself has been questioned by several researchers in recent years [8, 9, 12, 20, 25], leading to an effort for evaluating and rethinking metrics that provide a better alignment with the human-perspective.

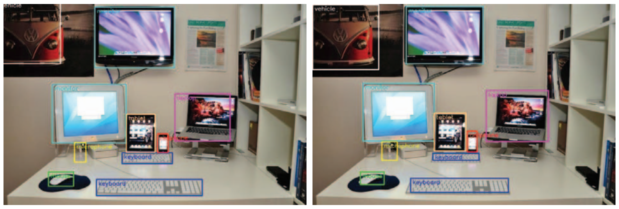


Figure 1. Discrepancies of bounding boxes (BBs) between two annotators.

A closely related problem is the lack of annotation consistency in existing datasets for object detection. This problem is inherent to the subjectivity of the annotation task, which relies on human annotators. For instance, many big datasets such as OID [11] and DOTA [30] present missing annotations due to human fatigue. Also, datasets often contain erroneous labels: a recent work [23] identified at least 34, 96, 50, and 23 label errors for BDD, Kitti, COCO, and VOC datasets, respectively. Finally, the process of annotating bounding boxes (HBBs – horizontal or OBB – oriented) is subject to fatigue or even annotator biases. For example, Figure 1 shows the bounding boxes produced by two different human annotators, which were requested to draw bounding boxes from eight classes (keyboard, mouse, monitor, laptop, tablet, cellphone, microphone, vehicle) of an image from the Objects365 dataset [24]. We can note that some discrepancies might be due to a lack of proper care (see laptop and keyboard boxes), but others indicate an intrinsic ambiguity in the process, as the vehicle in the top-left: should the rear mirror be included in the annotation (and add a significant portion of the background) or not?

Annotation discrepancies affect the quality metrics used to benchmark object detectors. Absent or wrongly labeled annotations impact the AP since they generate misleading false positives or negatives. However, imprecise (noisy) bounding box annotations must also be considered: as shown by several authors [15, 17, 18], even subtle annotation discrepancies might impact the IoU and hence the AP.

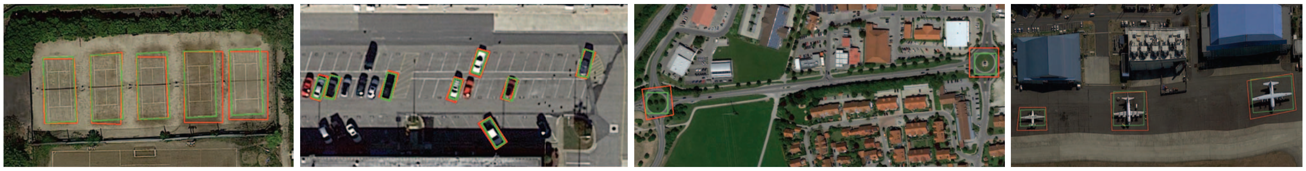


Figure 2. Discrepancies of oriented bounding boxes (OBB) based on DOTAv1.0 (red) and DOTAv1.5 (green).

One potential solution to mitigate annotation discrepancies is to hire multiple evaluators and integrate their annotations with a voting mechanism. However, this approach introduces a dilemma in determining the definitive bounding box parameters, and involves considerable costs in terms of money and time. While some studies [14, 32] aim to address noisy annotations, they assume noise levels that might deviate significantly from typical human errors: they introduced a uniform error between 10% to 40% of the original object size on COCO2017 [13] and VOC2007 [5] datasets.

The main goal of this work is to tackle *noisy* annotations, which correctly capture the object locations and the category labels but might contain *imprecision* in the bounding box. We aim to quantify this noise, evaluate its impact on the evaluation metrics, and propose alternatives that account for annotation imprecision. Although we focus on HBBs (HBBs will be referred to as BBs from this point on), the same problem happens – and might be even amplified – in OBB or 3D object detection. For example, Figure 2 illustrates some OBB annotations of the same objects in two versions of the same dataset, namely DOTAv1.0 (red) and DOTAv.15 (green) [30]. Even objects with well-defined rectangular oriented shapes, such as tennis courts or small vehicles, present annotation noise due to fatigue or pixel variations that were improved in the latter version. Other objects, such as airplanes and roundabouts, exhibit even higher error rates due to ambiguous boundary definitions or orientation.

In summary, the main contributions of this work are: i) We derive upper and lower bounds for any similarity metric for comparing BBs that satisfy the triangle inequality as a function of annotation noise, focusing on the IoU; ii) based on a set of original and corrected annotations for the same dataset (Microsoft COCO), we quantify the corner-wise annotation noise; iii) for a given corner-wise annotation noise distribution, we define noise-aware IoU metric that yields bounds for computing the AP; iv) we evaluate recent object detectors using the proposed noise-aware metrics, and conclude that the AP can be considerably under- or overestimated due to annotation noise.

2. Related Work

The literature on object detection is vast, and a recent survey can be found in [34]. Next, we provide a critical analysis of existing work that deals with evaluation metrics and annotation inconsistencies in datasets.

Evaluation metrics: The most popular evaluation metric for object detection is the mean Average Precision (AP) [34]. A key issue when computing the AP is the definition of a *correct* detection, which involves comparing the Intersection-over-Union (IoU) between the prediction and the ground truth (GT), and comparing it with a pre-defined threshold T. In the VOC2007 protocol [5], a single value $T = 0.5$ is used, whereas the COCO protocol [13] suggests the average of several thresholds $T \in \{0.5, 0.55, ..., 0.95\}$. Some datasets propose class-related IoU threshold values, such as the evaluation protocol in the KITTI dataset [6][1]. They use a 0.7 overlap for cars, and 0.5 for pedestrians and bicycles in both birds-eye and 3D object detection challenges.

Despite being widely (and sometimes blindly) used, both IoU and the AP have received several criticisms in the past years. Some authors mention that choosing the IoU threshold is arbitrary in the context of object detection or image segmentation [8, 20]. Other authors advocate the use of instance-dependent thresholds. For instance, Jeune and Mokraoui [9] propose an adaptive scale IoU to better discriminate small objects in the context of few-shot object detection.

The use of the IoU itself as a metric has also been questioned by several authors. Strafforello and colleagues [25] argue that humans go beyond a simple IoU number when visually evaluating object detectors. Another limitation of the IoU relates to non-overlapping objects, for which the IoU is zero, regardless of whether the objects are close or far apart. The Generalized IoU (GIoU) [22] mitigates the non-overlapping issue by also considering bounding boxes that encompass the two objects. The *subset case* (i.e., when one object is inside another) is another known limitation of the IoU, which yields the same value regardless of the relative location. The Probabilistic IoU (ProbIoU) [19] was recently introduced as an alternative metric to the IoU for both HBB and OBB detectors, mitigating both the subset and non-overlapping issues. Finally, some limitations of the AP metric have also been questioned recently: Jena and colleagues [8] noted that significant gains in AP can be achieved even if several false positives are introduced in the high-recall range.

Inconsistencies in datasets: Several authors studied the

[1] http://www.cvlibs.net/datasets/kitti/

effect of dataset annotation inconsistencies in the trained models. Papadopoulos *et al*. [21] introduced the "extreme clicking" annotation strategy for BBs and noted an average IoU of 0.88 compared to VOC 2017 GT annotations. An experiment comparing 50,000 boxes annotated by different humans was reported in [11], with a very similar result of 0.87 IoU. Murrugarra *et al*. [18] pointed out that annotation noise is more harmful w.r.t. the IoU for smaller objects, which has also been noted in [3, 26] when benchmarking small object detection. Jiaxin *et al*. [17] identified several annotation errors in the popular datasets COCO [13] and OID [11], and provided corrected annotations. They show that training detectors with such a fix yields significant improvements in both datasets. In contrast, Agnew *et al*. [1] aimed to empirically relate the effect of annotation noise on the AP by training object detectors with noisy bounding boxes. They noted an AP degradation of 0.185 when the largest amount of uniform noise was injected.

Another recent trend is to propose noise-aware object detectors, which should inherently deal with imprecision. For instance, DISCO [32] analyzes the distribution of proposal boxes to overcome noise. The methods proposed in [15, 29] treat detection as a multiple instance learning (MIL) approach where the classifier aims to filter inaccurate bounding boxes. Nevertheless, these methods introduce non-human or exaggerated errors, which may obscure their conclusions regarding annotation errors present in current datasets.

In this work, we propose a generic framework for noise-aware evaluation of object detection that provides upper and lower bounds for a baseline similarity metric. The framework can be applied to different object representations (2D or 3D HBBs, or OBBs) and any baseline similarity metric that induces a mathematical distance metric satisfying the triangle inequality, such as IoU, GIoU, or ProbIoU. We focus our analysis on 2D HBB object detection using the IoU, and evaluate recent object detectors using the proposed metrics.

3. The proposed approach

This section proposes an alternative evaluation metric for comparing bounding boxes that account for annotation uncertainty. We first discuss the importance of having actual mathematical metrics in the evaluation process and how they can be used to estimate error bounds. We then focus the analysis on the IoU, and how corner-wise annotation noise impacts the IoU computation. We explore a set of corrected bounding box annotations for the COCO dataset provided in [17] to estimate how annotation noise behaves in practice, and how such noise impacts the IoU. Finally, we propose noise-aware versions for the IoU based on these estimates.

3.1. Bounds for evaluation metric errors

As noted by Nguyen and colleagues [20], one of the requirements for designing a trustworthy evaluation metric is having consistency with mathematical requirements, such as the metric properties. In particular, the triangle inequality is a required property for a distance metric. For a given space $\mathcal{S}$, a distance metric $d : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ must satisfy

$$d(x, z) \leq d(x, y) + d(y, z), \forall x, y, z \in \mathcal{S}. \qquad (1)$$

As an immediate consequence of the triangle inequality, we have that

$$d(x, z) \geq d(x, y) - d(y, z), \forall x, y, z \in \mathcal{S}. \qquad (2)$$

If $\mathcal{S}$ is the space of bounding boxes, there are pairs of similarity/distance metrics that satisfy the triangle inequality. For example, $d_{\text{IoU}} = 1 - \text{IoU}$, where $\text{IoU} \in [0, 1]$, $d_{\text{GIoU}} = \frac{1-\text{GIoU}}{2}$, where $\text{GIoU} \in [-1, 1]$ [22], and $d_{\text{ProbIoU}} = 1 - \text{ProbIoU}$, where $\text{ProbIoU} \in [0, 1]$ [19]. We will proceed with our analysis with the IoU, which is the most popular metric, but analogous reasoning can be applied to ProbIoU and GIoU.

Let us consider that $\texttt{Det} \in \mathcal{S}$ is a *prediction* produced by an object detector, $\texttt{Ann} \in \mathcal{S}$ is a possibly *noisy annotation*, and $\texttt{GT} \in \mathcal{S}$ is the *ground truth (GT)*. The following relationships are immediate from Inequalities (1) and (2):

$$\text{IoU}(\texttt{Det}, \texttt{GT}) \geq \underbrace{\text{IoU}(\texttt{Det}, \texttt{Ann})}_{\text{Observed}} - \underbrace{d_{\text{IoU}}(\texttt{Ann}, \texttt{GT})}_{\text{Noise}}, \qquad (3)$$

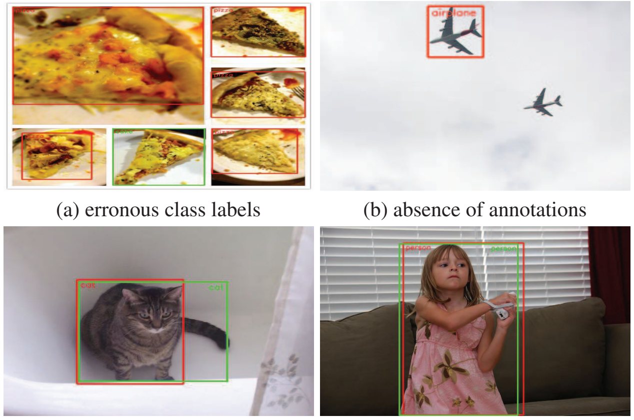
$$\text{IoU}(\texttt{Det}, \texttt{GT}) \leq \underbrace{\text{IoU}(\texttt{Det}, \texttt{Ann})}_{\text{Observed}} + \underbrace{d_{\text{IoU}}(\texttt{Ann.GT})}_{\text{Noise}}. \qquad (4)$$

The RHS of Inequalites (3) and (4) provide lower and upper bounds, respectively, for the *actual* similarity $\text{IoU}(\texttt{Det}, \texttt{GT})$ between the prediction and the GT considering the *observed* similarity $\text{IoU}(\texttt{Det}, \texttt{Ann})$ and the *annotation noise* distance $d_{\text{IoU}}(\texttt{Ann}, \texttt{GT})$. Since the latter term is typically unknown, it must be estimated.

3.2. Annotation Noise vs. IoU

A key issue in Inequalites (3) and (4) is to provide an estimate for $d_{\text{IoU}}(\texttt{Ann}, \texttt{GT})$. We follow the idea presented in [18] and provide a probabilistic relationship between the annotation noise of the top-left and bottom-right coordinates and the corresponding IoU value.

Without loss of generality, let us consider a canonical GT bounding box with top-left coordinates $\boldsymbol{x}_{tl}^{gt} = (0, 0)$ and bottom-right coordinates $\boldsymbol{x}_{br}^{gt} = (W, H)$, where W and H are the width and height, respectively. The coordinates of a related noisy annotation are given by $\boldsymbol{x}_{tl}^{ann} = (x_1, y_1)$ and $\boldsymbol{x}_{br}^{ann} = (W + x_2, H + y_2)$, where $\boldsymbol{\eta} = (x_1, x_2, y_1, y_2)$ is the corner-wise annotation noise. The intersection and union of the two boxes can be written as a function of the GT box



(a) erroneous class labels (b) absence of annotations (c) different interpretations (d) bounding box imprecision

Figure 3. Different types of errors/ discrepancies among annotators.

size $\boldsymbol{s} = (W, H)$ and the annotation noise $\boldsymbol{\eta}$ as

$$\begin{aligned} \mathrm{I}_{\boldsymbol{s}}(\boldsymbol{\eta}) &= (W + x_2^- - x_1^+)(H + y_2^- - y_1^+), \\ \mathrm{U}_{\boldsymbol{s}}(\boldsymbol{\eta}) &= HW + (W + x_2 - x_1)(H + y_2 - y_1) - \mathrm{I}_{\boldsymbol{s}}(\boldsymbol{\eta}), \end{aligned} \tag{5}$$

where $x^+ = \max\{x, 0\}$ and $x^- = \min\{x, 0\}$. We assume that noise is small and $|x_1|, |x_2| \leq W/2$, $|y_1|, |y_2| \leq H/2$. Hence, the corresponding IoU is given by

$$\mathrm{IoU}_{\boldsymbol{s}}(\boldsymbol{\eta}) = \frac{I_{\boldsymbol{s}}(\boldsymbol{\eta})}{U_{\boldsymbol{s}}(\boldsymbol{\eta})}. \tag{6}$$

Since the noise parameters $\boldsymbol{\eta}$ are unknown for a given annotation, we follow a probabilistic approach. More precisely, we compute the expected value $\overline{\mathrm{IoU}}_{\boldsymbol{s}} = \mathbb{E}[\mathrm{IoU}_{\boldsymbol{s}}]$ as

$$\overline{\mathrm{IoU}}_{\boldsymbol{s}} = \int_{\boldsymbol{\eta}} \mathrm{IoU}_{\boldsymbol{s}}(\boldsymbol{\eta}) p(\boldsymbol{\eta}) d\boldsymbol{\eta}, \tag{7}$$

where p is PDFs of the random variable $\boldsymbol{\eta}$.

If we know the distribution $p(\boldsymbol{\eta})$, we can compute the expected value for each bounding box dimension $\boldsymbol{s} = (W, H)$. Hence, we can find an estimate for the expected value of $\overline{d_{\mathrm{IoU}}(\mathtt{Ann}, \mathtt{GT})}$, which *depends on the dimensions* of the GT bounding box.

3.3. Estimating the annotation noise

As mentioned in Section 1, errors or discrepancies among annotators might relate to different class labels, absence of annotations, different interpretations of where exactly the boundaries of the objects are, or bounding box noise, as illustrated in Figure 3. This work focuses on the effect of bounding box noise, as shown in Figure 3d.

To empirically determine the behavior of annotation noise caused by human error, we need sets of images with bounding boxes annotated by two or more humans. Although we are not aware of any publicly available dataset with these characteristics, we can explore efforts that try to refine existing datasets for object detection. For instance, Ma *et al*. [17] provided “corrected” annotations for a subset of categories in two popular datasets: Microsoft COCO [13] and Open Images Dataset (OID) v4 [11], with a much larger number of reannotated BBs for COCO (569,309 vs. 24,995 for OID v4). Hence, we restrict our analysis to the COCO dataset.

Since the corrections also include mislabeled classes and missing objects, we present an approach for automatically filtering only bounding box *localization* corrections. We initially perform a per-image alignment of the original bounding box annotations and the corrected versions, which we assume to be the actual GT. Since the number of annotations per image might be different, we use the Hungarian algorithm [10] to perform bipartite graph matching using d_{IoU} as the cost function. In the original annotations of COCO, a single bounding box was used to annotate several instances of the same category for some images, whereas the corrected annotations pinpoint each instance, as illustrated in Figure 4. Hence, the Hungarian algorithm might still provide wrong alignments, as shown in the middle of Figure 4. To further refine the alignment step, we remove pairs of BBs that present IoU smaller than a threshold (empirically set to 0.1), as illustrated in the right of Figure 4. This process leads to a total of 284,392 paired BBs.

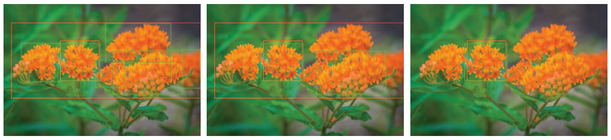


Figure 4. Pairing original (red) and corrected (green) annotations. Left: raw images with all annotations. Middle: pairing after Hungarian algorithm. Right: filtering wrong matches based on IoU threshold.

With the set of paired BBs, we compute the top-left (x_1, y_1) and bottom-right (x_2, y_2) annotation discrepancies, yielding four distributions – one for each coordinate. Since these values originated from annotation *imprecision*, we assume that they follow i.i.d. distributions according to a single individual PDF p_i, so that $p(\boldsymbol{\eta}) = p_i(x_1)p_i(x_2)p_i(y_1)p_i(y_2)$.

To estimate p_i, we initially consider all the samples x_1, y_1, x_2, y_2 obtained from the BB pairing procedure. We noticed a very long-tailed distribution, with some values around -500 or +500. Considering that the maximum dimension (width or height) in the paired dataset is 640, such large noise values are clearly outliers. To further refine the dataset with paired BBs, we used Tukey’s fence [27] based on interquartile distances to remove outliers. A value

$z \in \{x_1, x_2, y_1, y_2\}$ is considered an outlier if

$$z < Q_1 - 1.5\text{IQR} \;\text{ or }\; z > Q_3 + 1.5\text{IQR}, \tag{8}$$

where Q_1 and Q_3 are the first and third quantiles of the distribution, respectively, and $\text{IQR} = Q_3 - Q_1$ is the interquartile range. A pair of BBs is kept if *all* the four noise estimates x_1, y_1, x_2, y_2 are considered *inliers* by Tukey's fence, which leads to a total of 165,840 pairs of BBs. The average pair-wise IoU of the refined set is 0.89, which is very close to the results comparing human annotation discrepancies reported in [11, 21], and more information regarding the refined dataset is provided in the supplementary material.

The histogram of the resulting noise estimates for the refined dataset with paired BBs is shown at the top of Figure 5. We note a sharp peak at the origin and an apparently exponential decay along the tails, which leads us to choose a Laplace distribution as a parametric approximation. More precisely, we adopt a zero-mean truncated Laplace distribution $p_i(x)$ whose PDF is given by

$$p_i(x) = p(x; \sigma, M) = \frac{e^{-\frac{x-M}{\sigma}}}{2\sigma\left(e^{\frac{M}{\sigma}} - 1\right)}, \tag{9}$$

where $[-M, M]$ is the support of the distribution and σ is the scale parameter. The fitted PDF is shown as an orange line at the top of Figure 5, providing a good representation of the empirical histogram. Note that the sharp peak at the origin might be due to unchanged annotations in [17] and could be removed when fitting the PDF.

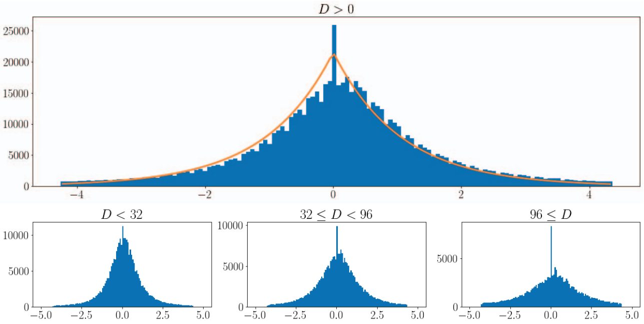


Figure 5. Histogram of noise bounding box (BB) components Top: overall distribution with all BBs. Bottom: individual histograms based on the size of the BB dimension D.

We used a single distribution to model the behavior of the corner-wise annotation noise $\boldsymbol{\eta} = (x_1, y_1, x_2, y_2)$ regardless of the BB dimension. On the other hand, previous works that explore annotation noise [15, 29] assumed a uniform perturbation *proportional* to the BB dimension. To better assess the dependency of the annotation noise as a function of the BB size, we have grouped the coordinate-wise noise values for each dimension (x_1, x_2 for width and y_1, y_2 for height) based on their size following the definition of small (S), medium (M) and large (L) boxes of the COCO evaluation protocol. More precisely, we consider that a dimension $D \in \{H, W\}$ is small if $D < 32$, medium if $32 \leq D < 96$, and large if $D \geq 96$. For each of these three groups, we computed the histogram of the noise values, shown at the bottom of Figure 5. The plots indicate that the dimension size has some impact on the noise distribution, but it is far from being uniform as suggested in [15, 29].

Finally, we sample integer values of $\boldsymbol{s} \in [1, 640]^2$, which is the range of BB dimensions in COCO. Then, estimate $\overline{\text{IoU}}_{\boldsymbol{s}}$ by numerically integrating Eq. (7) using $p(\boldsymbol{\eta}) = p_i(x_1)p_i(x_2)p_i(y_1)p_i(y_2)$, where p_i is the fitted truncated Laplacian PDF. For sizes that were not sampled, we used bilinear interpolation.

3.4. Noise-aware IoU

As shown by Inequalities (3) and (4), the IoU value of a predicted BB might be either over or underestimated due to annotation noise. In any case, the difference between the *actual* and the *observed* IoU values is bounded by $d_{\text{IoU}}(\texttt{Ann}, \texttt{GT}) = 1 - \text{IoU}_{\boldsymbol{s}}(\boldsymbol{\eta})$, where $\boldsymbol{s} = (W, H)$ is the size of the GT bounding and $\boldsymbol{\eta} = (x_1, x_2, y_1, y_2)$ are the corner-wise annotation error.

Although it is impossible to know $\text{IoU}_{\boldsymbol{s}}(\boldsymbol{\eta})$ for each GT bounding box, we can estimate the expected value based on empirical or theoretical models for the corner-wise annotation errors. Since the values depend on H and W, an empirical estimation would require many samples *for each combination* of height and width, which is unfeasible. On the other hand, the analysis provided in the previous sections allows us to estimate the expected value *for any combination* of height and width by using a parametric model for the noise distribution, such that $\mathbb{E}[d_{\text{IoU}}(\texttt{Ann}, \texttt{GT})] = 1 - \mathbb{E}[\text{IoU}_{\boldsymbol{s}}(\boldsymbol{\eta})] = \bar{\epsilon}(\boldsymbol{s}) = \bar{\epsilon}(W, H)$.

Table 1 shows a comparison between the empirical and theoretical estimation of $\mathbb{E}[\text{IoU}_{\boldsymbol{s}}(\boldsymbol{\eta})]$. For the empirical estimations, GT BBs were grouped into nine 2D bins based on the individual dimensions H and W, with three 1D bins for each dimension (small, medium, and large, as done in Figure 5). The theoretical estimation for each 2D bin was computed by averaging $\overline{\text{IoU}}_{\boldsymbol{s}}$ for all values $\boldsymbol{s} = (W, H)$ within the bin. The smallest IoU values were obtained when both H and W are small, and the theoretical estimate is a little lower ($\sim 5\%$) than the empirical estimate. For the remaining bins, the theoretical and empirical estimations are very close, with discrepancies around 1% or smaller.

For a detection $\texttt{Det}$ and annotation $\texttt{Ann}$ with dimension $H \times W$, the expected upper and lower IoU values between Det and the corresponding noise-free $\texttt{GT}$ are given by

$$\begin{aligned} \overline{\text{IoU}}^{u}(\texttt{Det}, \texttt{GT}) &= \text{IoU}(\texttt{Det}, \texttt{Ann}) + \bar{\epsilon}(H, W), \\ \overline{\text{IoU}}^{l}(\texttt{Det}, \texttt{GT}) &= \text{IoU}(\texttt{Det}, \texttt{Ann}) - \bar{\epsilon}(H, W), \end{aligned} \tag{10}$$

H/W	S	M	L
S	0.8048/0.7631	0.8732/0.8667	0.9062/0.9116
M	0.8850/0.8742	0.9214/0.9280	0.9433/0.9534
L	0.9097/0.9064	0.9464/0.9549	0.9711/0.9789

Table 1. Comparison of the empirical/theoretical estimation of $\mathbb{E}[\text{IoU}_{\boldsymbol{s}}]$ for small (S), medium (M) and large (L) values for the width W and height H.

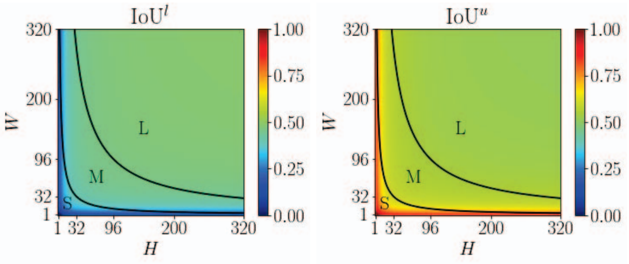


Figure 6. Noise-aware IoU values considering a baseline IoU of 0.5 for different annotation sizes, along with bounding box regions for small (S), medium (M), and large (L) objects.

and both estimates are clipped in the range $[0, 1]$ to keep the valid range of IoU values. The noise-aware IoUs are then used to compute adjusted AP^u_T and AP^l_T metrics for evaluating an object detector, which provide upper and lower bounds for the baseline AP_T.

4. Noise-Aware Evaluation of Object Detectors

In this section, we re-evaluate several state-of-the-art object detectors using the proposed noise-aware metrics. We selected four detectors that have different checkpoints available in the MMDetection[2] framework [2], namely Reppoints [31], the YOLOv10 family [28], the RTMDET family [16] and CO-DETR [33] with different backbones. CO-DETR-SwinL* has the same backbone as CO-DETR-SwinL but was pre-trained with Objects365 dataset.

We evaluated these detectors using the COCO 2017 validation split using the traditional (baseline) AP_T metrics (Table 2) and the proposed upper- and lower-bounds AP^u_T and AP^l_T (Table 3). The results are also discriminated based on object size (small, medium, or large) and different IoU thresholds T. As expected, the largest discrepancies between the baseline and adjusted metrics are observed for restrictive thresholds (in particular, see the results for $T = 0.95$). This effect is aggravated for small objects, and we see that $\text{AP}^l_{95} = 0$ for small objects in all detectors because it is impossible to guarantee a 0.95 IoU for the estimated noise level. On the other hand, the differences between AP_T and the bounds AP^u_T, AP^l_T for larger objects and less restrictive thresholds T are considerably smaller, which corroborates the results in Table 1 and Figure 6. In fact, the AP gap between the upper and lower bounds, given by $\Delta\text{AP}_T = \text{AP}^u_T - \text{AP}^l_T$, can be viewed as an *uncertainty* estimate of the actual AP for an IoU threshold T. For instance, the average value for ΔAP_{50} considering all detectors is ~ 1.1, ~ 3.4, and ~ 13.5 for large, medium, and small objects, respectively. Hence, even for a conservative IoU threshold of 0.5, the AP estimates for small objects might be misleading. We also note that the rankings can change for some object sizes. For example, YOLOv10 S outperforms Reppoints X-101 FPN-DCN in $\text{AP}_{50:95}$ for small objects, but the opposite happens in $\text{AP}^u_{50:95}$.

Analyzing the consolidated AP results considering all object sizes and IoU thresholds (last column of the tables), we note that all detectors are more negatively affected by AP^l than positively affected by AP^u. For instance, the Yolov10 X AP_{50} for small objects is 55.6, and the corresponding AP^u_{50} and AP^l_{50} values are, 60.0 (+3.4) and 46.3 (-9.3). We observe that the gain in the upper bound is smaller than the loss in the lower bound, which might indicate that the detector produces instances with IoU above 0.5 with GT annotations that are disregarded by the lower bound, but not so many instances with IoU below 0.5 that would be considered by the upper bound. A similar behavior can be observed for other IoU thresholds T and detectors, particularly for small and medium objects.

Some visual results of object detection with Yolov10 X are shown in Figure 7. We can observe that several predictions look visually good and coherent with the annotation (in red) but can yield relatively low IoU values. For instance, the mouse in the middle of the fourth image presents an IoU smaller than 0.7, generating a false detection for more restrictive thresholds. Figure 8 shows the comparison of Yolov10 X with the original annotations (red) and the corrected ones (green) in COCO. In the first two images, $\text{IoU}(\texttt{Det}, \texttt{Ann}) > \text{IoU}(\texttt{Det}, \texttt{GT})$, meaning that the *observed* IoU is overestimated. On the other hand, the opposite happens for the last two images. Without knowing the *actual GT* annotation, which is typically the case, we cannot tell if the observed IoU is better or worse than the actual value, which highlights the importance of the proposed bounds. More visual results are provided in the supplementary material.

5. Discussion and Limitations

Discussion: Although we focused our analysis on HBB object detection using the IoU as the similarity metric, the formulation provided in Section 3.1 can be applied to generic object representations (OBBs, 3D HBBs or OBBs) and different similarity metrics that induce metric distances satisfying the triangle inequality, such as GIoU [22] or ProbIoU [19]. However, changing the object representation (2D or 3D OBBs) or evaluation metric requires a different model to relate corner-wise noise $\boldsymbol{\eta}$ with the chosen metric.

[2] https://github.com/open-mmlab/mmdetection

Detector	Small				Medium				Large				All
	AP$_{50}$	AP$_{75}$	AP$_{95}$	AP$_{50:95}$	AP$_{50}$	AP$_{75}$	AP$_{95}$	AP$_{50:95}$	AP$_{50}$	AP$_{75}$	AP$_{95}$	AP$_{50:95}$	AP$_{50:95}$
Reppoints R-50 FPN	35.6	20.4	0.5	20.4	62.8	44.4	1.6	41.0	68.7	53.8	3.8	49.0	37.0
Reppoints R-101 FPN	39.8	23.8	0.7	23.4	67.1	48.6	2.1	44.7	74.0	58.8	5.3	53.2	40.5
Reppoints R-101 FPN-DCN	42.4	25.5	0.5	25.1	70.0	52.2	2.4	47.1	76.7	63.3	7.9	57.0	42.9
Reppoints X-101 FPN-DCN	45.0	26.4	0.5	26.2	71.3	52.8	3.0	48.4	78.2	65.1	7.3	58.5	44.2
YOLOv10 N	31.2	19.9	0.5	18.9	60.2	46.7	5.1	42.4	69.1	59.4	14.7	54.6	38.5
YOLOv10 S	42.7	28.7	0.9	26.8	69.7	56.7	7.5	51.0	77.9	69.2	21.5	63.8	46.3
YOLOv10 M	51.8	36.2	2.5	33.8	74.5	63.2	9.6	56.5	80.3	71.9	25.5	66.9	51.1
YOLOv10 B	53.6	37.9	2.3	35.1	75.9	64.2	10.4	57.8	81.2	73.9	27.4	68.4	52.5
YOLOv10 L	54.0	38.6	2.6	35.8	76.6	65.1	10.6	58.5	82.3	74.7	28.5	69.3	53.1
YOLOv10 X	55.6	40.4	2.9	37.1	77.9	66.4	11.9	59.9	84.1	76.4	29.6	71.0	54.4
RTMDet-tiny	35.2	21.5	0.5	21.0	65.0	50.4	4.6	45.5	73.8	63.6	14.4	58.3	41.1
RTMDet-s	41.3	26.6	1.4	25.3	68.2	53.9	5.6	48.7	77.8	68.5	16.7	62.6	44.6
RTMDet-m	48.1	33.1	1.9	30.7	73.6	60.5	7.0	54.1	80.5	72.5	21.9	66.5	49.4
RTMDet-l	52.2	36.6	2.6	34.0	75.3	62.5	8.6	56.2	82.0	74.5	24.3	68.5	51.5
RTMDet-x	54.7	39.4	2.1	36.0	76.5	63.9	9.0	57.4	82.8	74.6	24.9	69.2	52.8
CO-DETR R-50	52.1	38.4	2.2	34.8	73.7	61.9	9.6	55.6	80.3	72.8	24.3	67.1	52.0
CO-DETR SwinL	62.7	46.8	3.0	42.6	81.1	69.9	12.5	62.7	88.7	80.6	31.5	75.1	58.9
CO-DETR SwinL*	70.5	55.3	5.1	49.9	84.9	75.2	15.4	67.6	90.5	84.2	35.7	78.4	64.1

Table 2. AP$_T$ values (%) for different object detectors in COCOval with the standard AP$_T$ metric.

Detector	Small				Medium				Large				All
	AP$^u_{50}$	AP$^u_{75}$	AP$^u_{95}$	AP$^u_{50:95}$	AP$^u_{50}$	AP$^u_{75}$	AP$^u_{95}$	AP$^u_{50:95}$	AP$^u_{50}$	AP$^u_{75}$	AP$^u_{95}$	AP$^u_{50:95}$	AP$^u_{50:95}$
Reppoints R-50 FPN	39.6	31.7	15.5	30.7 **(+10.3)**	64.7	50.9	14.5	48.0 **(+7.0)**	69.3	56.7	11.0	52.0 **(+3.0)**	43.7 **(+6.7)**
Reppoints R-101 FPN	44.5	36.4	18.1	35.0 **(+11.6)**	69.0	55.4	17.6	52.3 **(+7.6)**	74.6	60.9	14.0	56.7 **(+3.5)**	48.4 **(+7.9)**
Reppoints R-101 FPN-DCN	47.1	38.2	19.7	37.0 **(+11.9)**	71.9	58.2	19.5	55.1 **(+8.0)**	77.2	65.3	18.2	60.4 **(+3.4)**	51.1 **(+8.2)**
Reppoints X-101 FPN-DCN	49.7	40.3	20.5	39.1 **(+12.9)**	73.2	59.9	20.8	56.5 **(+8.1)**	78.7	67.1	19.8	62.1 **(+3.6)**	52.9 **(+8.7)**
YOLOv10 N	35.6	28.4	15.2	27.8 **(+8.9)**	61.9	52.0	21.1	48.9 **(+6.5)**	69.6	61.0	27.0	57.5 **(+2.9)**	44.6 **(+6.1)**
YOLOv10 S	47.3	40.0	23.1	38.5 **(+11.7)**	70.9	61.7	28.5	58.2 **(+7.2)**	78.3	70.5	35.0	66.6 **(+2.8)**	53.6 **(+7.3)**
YOLOv10 M	56.7	48.8	30.2	47.3 **(+13.5)**	75.8	67.6	34.3	64.0 **(+7.5)**	80.7	73.1	40.8	69.8 **(+2.9)**	59.3 **(+8.2)**
YOLOv10 B	57.9	50.3	31.6	49.0 **(+13.9)**	77.0	68.5	36.7	65.3 **(+7.5)**	81.6	75.0	42.6	71.4 **(+3.0)**	60.9 **(+8.4)**
YOLOv10 L	58.4	51.3	32.9	49.8 **(+14.0)**	77.7	69.5	37.0	66.0 **(+7.5)**	83.2	76.0	43.9	72.4 **(+3.1)**	61.5 **(+8.4)**
YOLOv10 X	60.0	53.0	34.5	51.4 **(+14.3)**	79.1	70.4	38.9	67.5 **(+7.6)**	84.4	77.3	45.0	73.9 **(+2.9)**	63.0 **(+8.6)**
RTMDet-tiny	40.1	31.7	15.7	30.8 **(+9.8)**	66.6	55.8	21.7	52.5 **(+7.0)**	74.1	65.4	27.2	61.3 **(+3.0)**	47.7 **(+6.6)**
RTMDet-s	46.1	37.8	19.5	36.5 **(+11.2)**	69.5	59.5	24.2	55.8 **(+7.1)**	78.2	70.2	30.9	65.8 **(+3.2)**	51.8 **(+7.2)**
RTMDet-m	53.3	44.9	25.9	43.6 **(+12.9)**	74.8	65.4	30.9	61.9 **(+7.8)**	80.9	74.2	36.3	69.6 **(+3.1)**	57.3 **(+7.9)**
RTMDet-l	57.2	48.8	30.1	47.5 **(+13.5)**	76.5	68.1	33.6	64.0 **(+7.8)**	82.4	75.7	39.6	71.5 **(+3.0)**	59.7 **(+8.2)**
RTMDet-x	59.3	51.6	32.6	50.2 **(+14.2)**	77.7	68.5	34.8	65.3 **(+7.9)**	83.0	76.3	40.7	72.3 **(+3.1)**	61.3 **(+8.5)**
CO-DETR R-50	56.3	49.6	33.5	48.4 **(+13.6)**	74.7	66.6	34.6	63.0 **(+7.4)**	80.9	73.9	39.9	70.2 **(+3.1)**	60.4 **(+8.4)**
CO-DETR SwinL	67.2	60.5	41.6	58.6 **(+16.0)**	82.3	74.4	40.7	70.6 **(+7.9)**	89.0	82.0	47.8	78.3 **(+3.2)**	68.4 **(+9.5)**
CO-DETR SwinL*	74.2	68.5	49.6	66.3 **(+16.4)**	85.9	79.5	47.1	75.7 **(+8.1)**	90.7	85.3	52.9	81.5 **(+3.1)**	73.8 **(+9.7)**

Detector	Small				Medium				Large				All
	AP$^l_{50}$	AP$^l_{75}$	AP$^l_{95}$	AP$^l_{50:95}$	AP$^l_{50}$	AP$^l_{75}$	AP$^l_{95}$	AP$^l_{50:95}$	AP$^l_{50}$	AP$^l_{75}$	AP$^l_{95}$	AP$^l_{50:95}$	AP$^l_{50:95}$
Reppoints R-50 FPN	27.2	6.1	0.0	10.7 **(-9.7)**	60.4	34.3	0.1	33.5 **(-7.5)**	68.1	50.9	0.8	45.8 **(-3.2)**	29.8 **(-7.2)**
Reppoints R-101 FPN	31.0	6.9	0.0	12.2 **(-11.2)**	64.6	38.5	0.0	36.7 **(-8.0)**	73.1	56.0	0.1	49.8 **(-3.4)**	32.3 **(-8.2)**
Reppoints R-101 FPN-DCN	32.4	9.4	0.0	13.4 **(-11.7)**	66.9	41.8	0.0	38.8 **(-8.3)**	76.0	60.1	1.2	53.3 **(-3.7)**	34.2 **(-8.7)**
Reppoints X-101 FPN-DCN	33.7	8.5	0.0	13.6 **(-12.6)**	68.7	42.2	0.1	39.7 **(-8.7)**	77.2	62.5	1.2	54.8 **(-3.7)**	35.1 **(-9.1)**
YOLOv10 N	24.7	7.2	0.0	10.4 **(-8.5)**	58.3	38.8	0.2	35.6 **(-6.8)**	68.5	57.8	4.9	51.6 **(-3.0)**	32.2 **(-6.3)**
YOLOv10 S	35.1	10.1	0.0	14.9 **(-11.9)**	67.6	48.0	0.3	43.0 **(-8.0)**	77.4	67.5	9.1	60.4 **(-3.4)**	38.1 **(-8.2)**
YOLOv10 M	42.7	14.8	0.0	19.1 **(-14.7)**	72.8	54.4	0.5	47.6 **(-8.9)**	79.7	70.3	11.2	63.4 **(-3.5)**	41.7 **(-9.4)**
YOLOv10 B	43.9	15.9	0.0	19.6 **(-15.5)**	73.9	56.2	0.7	48.9 **(-8.9)**	80.7	72.6	12.1	65.0 **(-3.4)**	42.7 **(-9.8)**
YOLOv10 L	44.9	15.9	0.0	20.2 **(-15.6)**	74.9	56.4	0.6	49.4 **(-9.1)**	81.6	73.6	12.8	65.8 **(-3.5)**	43.3 **(-9.8)**
YOLOv10 X	46.3	17.0	0.0	21.0 **(-16.1)**	76.0	58.3	0.7	50.6 **(-9.3)**	82.9	75.2	13.0	67.2 **(-3.8)**	44.3 **(-10.1)**
RTMDet-tiny	28.2	7.5	0.0	11.7 **(-9.3)**	62.9	40.9	0.2	37.9 **(-7.6)**	73.3	61.6	5.0	55.0 **(-3.3)**	34.2 **(-6.9)**
RTMDet-s	33.8	9.3	0.0	14.1 **(-11.2)**	66.3	44.5	0.2	40.7 **(-8.0)**	77.4	66.9	6.0	59.1 **(-3.5)**	37.0 **(-7.6)**
RTMDet-m	39.9	12.3	0.0	17.2 **(-13.5)**	71.8	51.1	0.2	45.4 **(-8.7)**	80.1	70.1	8.4	62.8 **(-3.7)**	40.6 **(-8.8)**
RTMDet-l	43.6	15.7	0.0	19.3 **(-14.7)**	73.6	53.0	0.3	47.2 **(-9.0)**	81.5	73.3	9.2	64.8 **(-3.7)**	42.1 **(-9.4)**
RTMDet-x	45.6	14.5	0.0	20.3 **(-15.7)**	74.7	55.5	0.3	48.3 **(-9.1)**	82.4	73.3	10.5	65.5 **(-3.7)**	43.1 **(-9.7)**
CO-DETR R-50	42.9	15.3	0.0	19.1 **(-15.7)**	71.8	52.6	0.3	46.5 **(-9.1)**	79.9	71.2	10.4	63.6 **(-3.5)**	41.8 **(-10.2)**
CO-DETR SwinL	52.5	18.3	0.0	23.4 **(-19.2)**	79.5	61.1	0.4	52.8 **(-9.9)**	87.7	79.4	13.8	71.2 **(-3.9)**	46.9 **(-12.0)**
CO-DETR SwinL*	59.5	22.6	0.0	27.7 **(-22.2)**	83.7	66.8	0.6	57.1 **(-10.5)**	89.7	83.0	16.3	74.3 **(-4.1)**	50.7 **(-13.4)**

Table 3. AP^u_T and AP^l_T values (%) for different object detectors in COCOval, related to the IoU upper and lower bound metrics.

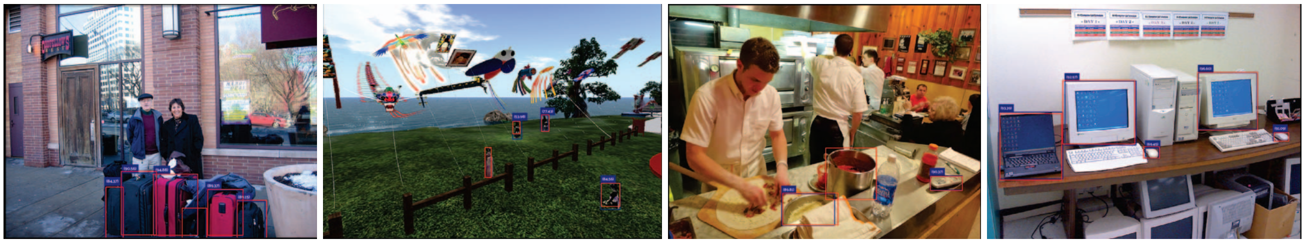


Figure 7. Detections results of Yolov10 X (blue) and annotations (red), along with the corresponding IoU values (in %). Best seen zoomed.

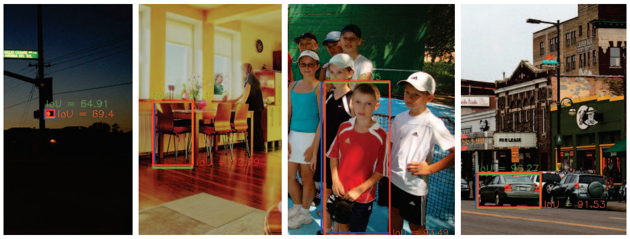


Figure 8. Image crops with detections results of Yolov10 X (blue), original (red) and corrected annotations (green) with their corresponding IoU (in %).

The empirical comparison between original and reannotated BBs performed in Section 3.3 yields an average IoU of 0.89, which is very close to inter-annotator experiments reported in [11, 21]. Our analysis presented in Section 4 is also aligned with the findings in [20], which show that object detection evaluation based on AP can be strongly affected by annotation noise at higher IoU thresholds. However, we show that the IoU degradation is strongly related to the BB dimensions, as shown in Table 1 and Figure 6. In particular, smaller objects are more sensitive to annotation noise, which can severely over- or under-estimate the traditional AP_T metric, particularly for more restrictive thresholds. We hope that the proposed noise-aware metrics, which are adjusted based on the BB dimensions and expected noise level, can provide a more comprehensive evaluation of object detectors.

Limitations: Eq. (7) can be used to estimate the expected IoU degradation caused by corner-wise perturbations assuming a generic PDF, and several assumptions/simplifications were assumed in this work. We used a single truncated Laplacian distribution as the PDF for every HBB, but Figure 5 indicates some variation depending on the bounding box dimensions. Also, we used the noise distribution from only five categories to formulate the class-agnostic metrics. Although the findings of Kuznetsova *et al*. [11] that imprecise boxes are quite evenly spread over classes corroborate our assumption, further studies might be needed. We also assumed that the reannotated boxes for Open Images provided by [17] reflect the GT boxes, but they might also contain noise. Finally, it would be interesting to evaluate how the noise distribution behaves for different datasets, i.e., if we can use the noise-aware metrics estimated from COCO to evaluate a different dataset.

6. Conclusions and Future Work

This paper presented a noise-aware framework for evaluating object detectors. From a baseline similarity metric that induces a mathematical distance metric, we deduced upper and lower bounds based on annotation noise. For HBB object detection and the IoU as the similarity metric, we relate the corner-wise noise level with the IoU degradation, showing that it can be significantly under- or over-estimated. In particular, smaller objects can be strongly affected by noise. Based on a reannotated version of the COCO dataset provided in [17], we estimated the corner-wise noise distribution and computed the expected IoU degradation for a bounding box based on its dimensions $H \times W$. Then, we introduced noise-aware expected upper and lower bounds for the IoU based on noise levels and BB dimensions, which are used to compute the corresponding AP_T upper and lower bounds (AP^u_T and AP^l_T). Finally, we evaluated state-of-the-art object detectors in the validation split of the COCO dataset using the traditional AP_T metrics and the proposed error-aware bounds, showing the results for different IoU thresholds T and object dimensions (small, medium, and large). As expected, the largest discrepancies between the baseline AP_T and the proposed bounds AP^u_T and AP^l_T were observed for small objects and more restrictive thresholds T. As an overall trend for all tested object detectors, the lower bound AP^l_T was more pessimistic than the upper bound AP^u_T was optimistic. In future work, we plan to explore the proposed bounds for defining object-aware thresholds instead of using a single value for all instances. Since smaller objects are more affected by noise, we can use more relaxed IoU thresholds compared to larger objects.

Acknowledgements: This study was partially funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We also want to acknowledge to LXAI supercomputer program, Gobierno de Jalisco, and Nvidia Corp for providing access to a DGX A-100 640 server.

References

[1] Cathaoir Agnew, Ciarán Eising, Patrick Denny, Anthony Scanlan, Pepijn Van De Ven, and Eoin M. Grua. Quantifying the effects of ground truth annotation quality on object detection and instance segmentation performance. *IEEE Access*, 11:25174–25188, 2023.

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[3] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13467–13488, 2023.

[4] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[5] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 06 2010.

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[7] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[8] Rohit Jena, Lukas Zhornyak, Nehal Doiphode, Pratik Chaudhari, Vivek Buch, James Gee, and Jianbo Shi. Beyond map: Towards better evaluation of instance segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11309–11318, 2023.

[9] Pierre Le Jeune and Anissa Mokraoui. Rethinking intersection over union for small object detection in few-shot regime. *arXiv preprint arXiv:2307.09562*, 2023.

[10] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[11] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[12] Yuxuan Li, Licheng Jiao, Zhongjian Huang, Xin Zhang, Ruohan Zhang, Xue Song, Chenxi Tian, Zixiao Zhang, Fang Liu, Shuyuan Yang, et al. Deep learning-based object tracking in satellite videos: A comprehensive survey with a new dataset. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):181–212, 2022.

[13] 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 *ECCV*, pages 740–755, 2014.

[14] Chengxin Liu, Kewei Wang, Hao Lu, Zhiguo Cao, and Ziming Zhang. Robust object detection with inaccurate bounding boxes. In *Proceeding of European Conference on Computer Vision (ECCV)*, 2022.

[15] Chengxin Liu, Kewei Wang, Hao Lu, Zhiguo Cao, and Ziming Zhang. Robust object detection with inaccurate bounding boxes. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 53–69, Cham, 2022. Springer Nature Switzerland.

[16] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors, 2022.

[17] Jiaxin Ma, Yoshitaka Ushiku, and Miori Sagara. The effect of improving annotation quality on object detection datasets: A preliminary study. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4849–4858, 2022.

[18] Jeffri Murrugarra-Llerena, Lucas N. Kirsten, and Claudio R. Jung. Can we trust bounding box annotations for object detection? In *CVPRW*, pages 4813–4822, 2022.

[19] Jeffri Murrugarra-Llerena, Lucas N. Kirsten, Luis Felipe Zeni, and Claudio R. Jung. Probabilistic intersection-over-union for training and evaluation of oriented object detectors. *IEEE Transactions on Image Processing*, 33:671–681, 2024.

[20] Tran Thien Dat Nguyen, Hamid Rezatofighi, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustworthy are performance evaluations for basic vision tasks? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8538–8552, 2023.

[21] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE international conference on computer vision*, pages 4930–4939, 2017.

[22] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019.

[23] Marius Schubert, Tobias Riedlinger, Karsten Kahl, Daniel Kröll, Sebastian Schoenen, Siniša Šegvić, and Matthias Rottmann. Identifying label errors in object detection datasets by loss inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4582–4591, January 2024.

[24] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019.

[25] Ombretta Strafforello, Vanathi Rajasekart, Osman S. Kayhan, Oana Inel, and Jan van Gemert. Humans disagree with the iou for measuring object detector localization error. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1261–1265, 2022.

[26] Kang Tong and Yiquan Wu. Rethinking pascal-voc and ms-coco dataset for small object detection. *Journal of Visual Communication and Image Representation*, 93:103830, 2023.

[27] John W Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

[28] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024.

[29] Di Wu, Pengfei Chen, Xuehui Yu, Guorong Li, Zhenjun Han, and Jianbin Jiao. Spatial self-distillation for object detection with inaccurate bounding boxes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6855–6865, 2023.

[30] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[31] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[32] Donghao Zhou, Jialin Li, Jinpeng Li, Jiancheng Huang, Qiang Nie, Yong Liu, Bin-Bin Gao, Qiong Wang, Pheng-Ann Heng, and Guangyong Chen. Disco: Distribution-aware calibration for object detection with noisy bounding boxes, 2024.

[33] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6725–6735, 2023.

[34] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.