Can we trust bounding box annotations for object detection?

Jeffri Murrugarra-Llerena LN Kirsten Claudio R. Jung
Institute of Informatics
Federal University of Rio Grande do Sul
{jeffri.mllerena, lnkirsten, crjung}@inf.ufrgs.br

Abstract

Object detection is a classical problem in computer vision, and the vast majority of approaches require large annotated datasets for training and evaluation purposes. The most popular representations are bounding boxes (BBs), usually defined as the minimal-area rectangle that encompasses the whole object region. However, the annotation process presents some subjectiveness (particularly when occlusions are present), and its quality might get degraded when the annotators get tired. Comparing BBs is crucial for evaluation purposes, and the Intersection-over-Union (IoU) is the standard similarity metric. In this paper, we provide theoretical and experimental results indicating that the IoU can be strongly affected even by small annotation discrepancies in popular datasets used for object detection. As a consequence, the Average Precision (AP) value commonly used to evaluate object detectors is also influenced by annotation bias or noise, particularly for small objects and tighter IoU thresholds.

1. Introduction

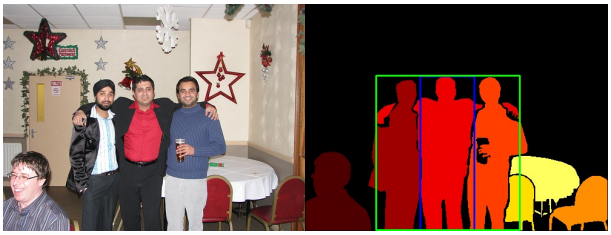
Object detection is a classical problem in computer vision and have greatly benefited from deep learning in the past years [12]. Training and evaluation datasets are growing larger and larger, and performing a thorough qualitative evaluation is unfeasible. As such, the development of quantitative performance metrics is crucial for evaluating and comparing different object detectors.

Training and evaluation of object detectors strongly rely on comparing the annotated object representation with the output of the detector. The most popular object representation is a bounding box (BB), which is the "smallest" rectangle that fully contains the object, where minimization is typically based on area. Within BB representations, axis-aligned rectangles – also called Horizontal Bounding Boxes (HBBs) – are the most common choices, and HBB annotations are present in popular datasets containing daily objects such as Pascal VOC 2012 [3] and Microsoft COCO 2017 [11]. HBBs are easy to annotate, and they require only four parameters, such as top-left and bottom-right coordinates. In the more generic case, Oriented Bounding Boxes (OBBs) also include a rotation angle. However, using OBBs involves additional challenges, such as the choice of a suitable parameterization that does not generate ambiguities, which is also related to the choice of the detector itself [30]. There has been an increasing interest in OBB object detection, and a few datasets provide OBB annotations, such as FDDB [8] (for face detection)[1], ICDAR 2015 [9], MLT 2017 [17] (for text spotting) and HRSC 2016 [14], DOTA (v1) [27] (aerial/satellite images).

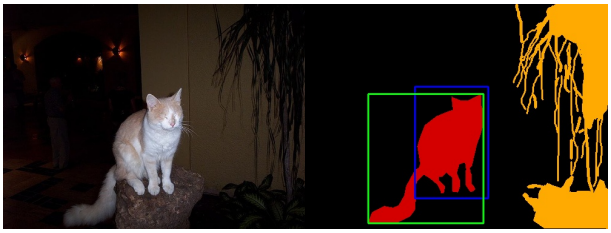
Although annotating HBBs or OBBs might seem easy and straightforward, it is not an entirely objective task. The definition of a BB itself for objects with strong partial occlusions is not trivial: should we annotate only the main visible portion or a BB comprising *all* object parts, regardless of their size? Furthermore, annotation tools typically involve drag-and-drop interfaces, which tend to generate annotation errors – hopefully of only a few pixels. To illustrate possible ambiguities or errors in annotated HBBs (AHBBs), we select some images that present both HBB annotations and segmentation masks, which allows us to compute the HBB estimated directly from the masks (SHBBs). Fig. 1 shows the RGB images and the corresponding segmentation masks with SHBBs overlaid in green and AHBBs in blue for a few samples in VOC 2012. In Fig. 1a, the AHBB comprehends both visible and "guessed" portions of the inner bike; the SHBB, on the other hand, relates only to the upper-central part of the bike. Fig. 1b illustrates a recurrent issue in VOC 2012: the object is partially occluded, and the segmentation mask presents more than one connected component with the body and hands, yielding a wide SHBB; the AHBB, however, relates only to the body. Fig. 1c shows an example where the AHBB annotator considered only the body of the cat with a small offset on the top-right corner, but the segmentation annotator also marked the tail. In all these examples, the IoU value between the AHBB and SHBB fell

[1] Annotations in FDDB are actually ellipses, which are mapped to OBBs.

(a) IoU = 0.47



(b) IoU = 0.38



(c) IoU = 0.38

(d) IoU = 0.74

Figure 1. Examples of HBB annotation ambiguities and discrepancies in VOC 2012.

below 0.5. Finally, Fig. 1d shows an example that seems to indicate human-center bias or annotation noise regarding the actual limits of the object.

Regardless of the object parametrization (HBBs, OBBs, or even full segmentation masks) and the annotation process, the Intersection-over-Union (IoU) [4] has been the *de facto* standard for comparing two shapes. A detection with region Ω_{det} is considered correct if $\text{IoU}(\Omega_{det}, \Omega_{gt}) \geq T$, where Ω_{gt} a GT region and $0 \leq T \leq 1$ is a threshold. Increasing the value of T forces a tighter match, and the value $T = 0.5$ was used in [4] for evaluating HBB object detectors. The standard in [11] is to compute several thresholds T varying from 0.5 to 0.95 in steps of 0.05, which provides overlap estimates with different tightness constraints. However, the choice of T is arbitrary, as discussed in [20]. In fact, a few evaluation protocols, such as the birds-eye-view object detection in the KITTI dataset [6][2] explores per-category thresholds: 0.7 for cars, and 0.5 for pedestrians and bicycles. Based on the IoU and the acceptance threshold T or a set of thresholds as in [11], the per-category *Average Precision* – computed based on precision and recall rates [4] – and the mean Average Precision (mAP) over all categories are commonly used as the overall performance metric.

In this paper, we perform a critical analysis to evaluate how BB uncertainties might impact the IoU and, consequently, the AP values that are used as default metrics to compare HBB and OBB object detectors. Since we are not aware of datasets containing object annotations of different humans, we perform our analysis by considering datasets that present both segmentation masks and BB annotations. Our results indicate that even sub-pixel discrepancies might considerably lower the IoU values, particularly for smaller objects. We hope our findings serve as basis for further research and critical analysis on the blind use of IoU and AP metrics for evaluating object detectors.

2. Related Work

The availability of annotated datasets is crucial for training and evaluating deep learning in a variety of tasks, ranging from image classification to object detection and instance segmentation. In most applications, the actual "ground-truth" itself is not objective, and datasets provide the view of one or several human annotators.

Even before the deep-learning boom, the generation and consistency check of human annotators was a concern. Martin et al. [15] proposed an image segmentation dataset annotated by different people. They noted that although the annotations were semantically consistent, human annotations for the same image vary in terms of granularity.

Misra and colleagues [16] evaluated the problem of "human-centric" annotations, in which the inherent subjectivity of the task affects the annotation process. They considered these labels noisy and proposed a deep network for decoupling the human reporting bias from the correct visual information in image tagging applications. Some papers evaluated bias and annotation errors – or "issues" – in large image classification datasets. Tsipras et al. [23] evaluate the annotation bias in large image classification datasets such as ImageNet [1], and how the creation process of the dataset might induce biases. They point out the presence of images with multiple valid labels and ambiguous classes, such as missile/projectile. Similar findings about class ambiguity were also reported in [18, 19] for ImageNet, which might affect quality metrics based on top-1 accuracy.

Rezatofighi et al. [20] presented a generic discussion on performance evaluation metrics for tasks such as object de-

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

tection and instance segmentation. They mention the drawback of IoU for disjoint regions, which is null regardless of their distance, and advocate for the use of the Generalized IoU [21] to overcome this issue. They also point out the effect of the IoU (or GIoU) acceptance threshold T to compute AP-related metrics, and show that order rankings of consolidated approaches for object detection and instance segmentation can change considerable when T changes.

Hall et al. [7] proposed a probabilistic representation of HBBs where the top-left and bottom-right corners are modeled as 2D Gaussian distributions, allowing the definition of a "probabilistic" HBB. A similar approach was presented in [25], where the authors also define a Jaccard IoU that compares two probabilistic boxes, which inherently accounts for the uncertainty of the compared HBBs.

Our works goes in the direction of [20] and evaluates the effect of IoU thresholds on the accuracy of object detectors. However, we focus on the "quality" of HBB annotations and how uncertainties impact the IoU value for popular datasets. We highlight the importance of the triangle inequality for an evaluation metric d as presented in [20]. If GT_r is the real (noise-free but unknown) ground-truth, GT_a is a (noisy or biased) ground-truth annotation and Det is a detected HBB, then

$$d(\text{Det}, \text{GT}_r) \leq d(\text{Det}, \text{GT}_a) + d(\text{GT}_r, \text{GT}_a). \quad (1)$$

If we have an estimate for the annotation noise/bias $d(\text{GT}_r, \text{GT}_a)$, then Eq. (1) provides an upper bound for the actual distance $d(\text{Det}, \text{GT}_r)$ based on the observed distance $d(\text{Det}, \text{GT}_a)$.

3. Evaluating dataset self-consistency for HBBs

The main concept behind the term *bounding box* is that the box should entirely contain the object under consideration. Although this concept might be clear for fully visible objects, partial occlusions either in the middle of the object or its extremities might generate annotation ambiguities. Since we are not aware of datasets containing HBB annotations from different humans to the same object, we "emulate" the human error considering datasets that provide both HBB annotations and segmentation masks.

COCO 2017: (will be called simply COCO from now on) provides both segmentation masks and AHBBs for a variety of objects in 80 different categories. As mentioned in [11], AHBBs are obtained directly from the polygonal regions that define segmentation masks, which are stored in a sub-pixel level and hence so does the segmentation-induced HBBs, called SHBB here. On the other hand, typical annotations for AHBBs are performed by humans that directly draw on the images, and usually consist of integer pixel coordinates, as in VOC [4].

In this first experiment, we "emulate" an *ideal* human annotator that is able to generate AHBBs with integer coordinates that best match the segmentation mask in COCO. More precisely, we rounded the floating-point SHBBs coordinates of COCO's training set using a `floor` operator for top-left coordinates and `ceil` for bottom-right to generate a *bounding* representation of the segmentation mask, and analyze the IoU between the SHBB and the AHBB.

Figure 2 shows a per-category boxplot of the IoU values, and we note that some categories are strongly affected by these sub-pixel changes, such as `car`, `traffic light`, `birds`, `sports ball` and `book`. This means that even if an object detector can accurately mimic the SHBB annotations, which is the available information in COCO, the AP values could be considerably degraded when compared to the AHBB or vice-versa.

Although we noted that some categories are more affected by HBB perturbations, the underlying reason is not the category itself, but the size of the HBBs. The round-off procedure generates per-coordinate absolute errors (x or y) smaller than 1 pixel, and the effect on the IoU is highly dependent on the dimensions of the HBB: if the width or height is small, even such a small error might considerably degrade the IoU. Figure 3 shows a category-agnostic scatter plot of the IoU vs. the smallest SHBB dimension for all annotations in COCO (blue points), which indicates that the IoU is related to the minimum SHBB dimension. In fact, the Spearman order rank correlation coefficient for the scatter plot is 0.9367 (with a numerically null p-value), which indicates a clear monotonic relationship between the smallest SHBB dimension and the IoU.

For the experiment with COCO, we can formally estimate the effect of coordinate rounding-off on the IoU values. Let us consider that an SHBB with dimensions $W \times H$ is the *ideal* GT annotation and that the AHBB obtained by the rounding-off procedure is the *best realizable* annotation with integer coordinates. Let us also consider that the horizontal (left and right) and vertical (top and bottom) offsets used to round-off the SHBB, given by $\boldsymbol{x} = (x_1, x_2, y_1, y_2)$, respectively, follow a uniform distribution $\mathcal{U}(0, 1)$. In this case, the IoU between the SHBB and an AHBB of offset $\boldsymbol{x}$ is given by

$$\text{IoU}_{\boldsymbol{x}}(W, H) = \frac{WH}{(W + x_1 + x_2)(H + y_1 + y_2)}, \quad (2)$$

and the expected IoU value is given by

$$E[\text{IoU}(W, H)] = \int_{\boldsymbol{x} \in R} \text{IoU}_{\boldsymbol{x}}(W, H) d\boldsymbol{x} = f(W) f(H), \quad (3)$$

where $R = [0, 1]^4$, and

$$f(x) = x \left(x \ln \left(\frac{x(x+2)}{(x+1)^2} \right) - 2 \ln \left(\frac{x+1}{x+2} \right) \right). \quad (4)$$

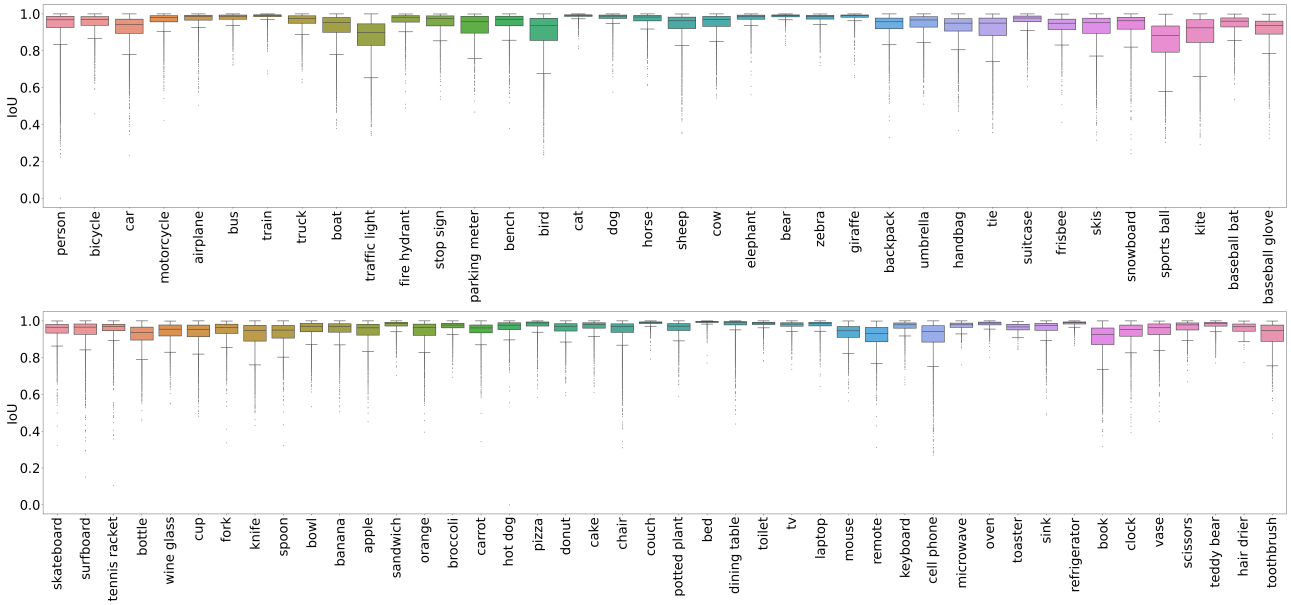


Figure 2. Per-category Iou between AHBBs and HBBs generated from segmentation masks (SHBBs) in COCO.

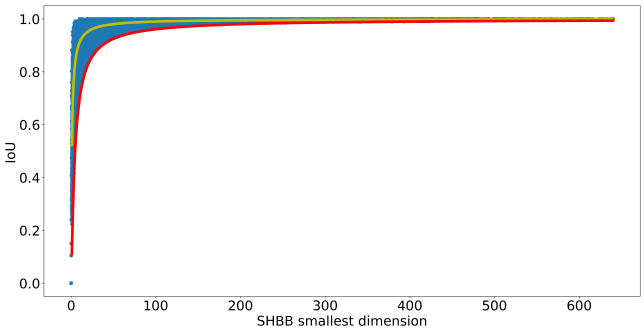


Figure 3. Scatter plot showing the IoU between HBB/SHBB pairs in COCO vs. the smallest SHBB dimension for COCO (blue). Also shows the theoretical lower bound (red) and the expected IoU value (yellow) considering only the smallest dimension according to Eq. (4)

We can observe that $E[\text{IoU}(W, H)]$ presents a separable contribution of the width W and height H, both guided by the same monotonically decaying function f. For the sake of illustration, the plot of f is shown as the yellow curve in Figure 3, overlaid with the scatter plot of the observed IoU values in the experiment. Note that f relates to the *expected* IoU value for a *single dimension*, but we can also compute a lower bound based on the smallest SHBB dimension. Given a minimum dimension $d = \min\{H, W\}$, Eq. (2) indicates that the minimum IoU value is achieved when $H = W$, (i.e., when the largest dimension equals the smallest) and $\boldsymbol{x} = (1, 1, 1, 1)$, which is the largest possible round-off error. In this case, the lower IoU bound is given by $d^2/(d+2)^2$, shown as a red curve in Figure 3. We can see that some SHBB/HBB pairs get very close to this lower bound.

As a final experiment with the COCO dataset, we evaluate how the annotation discrepancies can affect the AP metrics used to evaluate object detectors. For example, let us consider that the SHBBs are the *observed* annotations and AHBBs are the *actual* annotations for COCO or vice-versa. If a test image presents N objects, an *ideal* object detector would predict exactly N objects with their corresponding categories at a perfect detection score of one and would be able to regress the bounding box parameters according to the observed annotations. Using the validation set of COCO, for which we have GT annotations, we computed the AP_T values for such ideal detector with the 10 thresholds T suggested in [11] varying linearly from 0.50 to 0.95, and evaluated the results for *small*, *medium* and *large* objects, which are separated by area thresholds of 32^2 and 96^2 as in the official evaluation tools of COCO. We also performed a similar experiment using *real* detectors that were trained with the SHBB annotations in COCO and evaluated using both SHBB and AHBB annotations. Table 1 shows that the ideal detector is not affected by sup-pixel discrepancies at all IoU levels for large objects. However, for medium and particularly for small objects, there is a strong IoU decrease for larger values of T, which is consistent with Figure 3. For the experiments with real object detectors, we chose members of the EfficientDet [22][3] and YoloR [24][4]

[3]Code and weights from https://github.com/google/automl/tree/master/efficientdet

[4]Code and weights from https://github.com/WongKinYiu/

Detector	Small				Medium				Large			
	AP_{50}	AP_{75}	AP_{95}	$AP_{50:95}$	AP_{50}	AP_{75}	AP_{95}	$AP_{50:95}$	AP_{50}	AP_{75}	AP_{95}	$AP_{50:95}$
Ideal	100.0 / 99.60	100.0 / 93.46	100.0 / 13.63	100.0 / 82.38	100.0 / 100.0	100.0 / 100.0	100.0 / 77.15	100.0 / 97.66	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0
EfficientDet D0	24.87 / 25.34	11.54 / 11.62	0.20 / 0.15	12.53 / 12.86	61.01 / 61.01	41.58 / 41.85	1.52 / 1.55	38.61 / 38.76	72.60 / 72.67	58.37 / 58.34	6.27 / 6.52	52.59 / 52.72
EfficientDet D7	55.93 / 56.31	39.49 / 38.79	2.93 / 1.63	36.92 / 36.36	77.16 / 77.30	63.17 / 63.43	9.18 / 9.12	57.29 / 57.38	82.04 / 82.05	71.77 / 71.89	21.20 / 22.71	66.73 / 67.00
YOLO-R P6	56.02 / 56.58	40.73 / 39.22	2.48 / 1.60	37.04 / 36.22	76.02 / 76.12	63.10 / 63.00	9.34 / 8.32	56.85 / 56.52	79.95 / 79.97	72.08 / 72.31	21.92 / 21.51	65.96 / 65.94
YOLO-R W6	57.27 / 57.65	40.81 / 40.26	2.57 / 1.56	37.92 / 36.96	76.85 / 76.92	64.63 / 64.43	10.28 / 9.78	58.09 / 57.83	81.55 / 81.55	72.51 / 72.60	24.48 / 23.84	67.31 / 67.27

Table 1. AP_T values (%) for a different HBB object detectors in COCO trained with SHBBs and evaluated with SHBBs/AHBBs

models. We can observe that all detectors present small $AP_{50:95}$ (i.e., the mean of result from AP_{50} to AP_{95} varying the threshold in 5 units) variations when changing from SHBBs to AHBBs for large and even medium objects, but the detectors with the best results – Yolo-R P6 and Yolo-R W6 – suffer some degradation in AP_{95} for medium objects. For small objects, the AP_T values are more affected by the annotation format used for validating the results, in special for $T \in \{75, 95, 50:95\}$. In particular, Yolo-R W6 presents the best $AP_{50:95}$ results using SHBB annotations for small objects, but is only the third-best when considering AHBBs. It is also interesting to note that Efficiendet D0 presented higher $AP_{50:95}$ for small objects when using AHBBs. Our experiments with COCO show that small and medium objects are more susceptible to subpixel errors, as expected.

VOC 2012: presents HBB annotations (AHBBs) for 20 different categories, and a subset of 2,913 images also presents segmentation masks – we will refer to this subset simply as VOC from now on. In fact, the annotation process for VOC evolved in time, and more effort was put into generating and checking segmentation masks [2]. Hence, VOC is an interesting case study to check how consistent the HBBs generated automatically from the segmentation masks (SHBBs) are with the manually annotated AHBBs. Unlike the previous experiment with COCO, the AHBBs are not derived by rounding-off SHBBs, which might lead to discrepancies larger than one pixel.

There is no tagging between the segmentation mask and AHBB annotations in the dataset – in fact, the number of objects in both representations does not match for 6 of the 2,913 images. For the remaining 2,907 images, we adopted the following steps:

1. Compute the SHBB from each segmentation mask present in the image

2. Perform a pair-wise matching between SHBBs and AHBBs using the Hungarian algorithm [10], where the cost matrix was the negative IoU value

3. Discard matches for which the category annotations did not match (only two instances in the total)

4. Store the IoU values for each category in the dataset

yolor

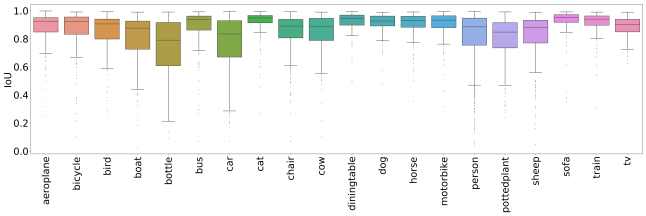


Figure 4. Iou between annotated HBBs (AHBBs) and HBBs generated from segmentation masks (SHBBs) for VOC.

At the end of this process, we obtained a set of 6,909 pairs of matched SHBBs and AHBBs, along with the corresponding IoU. The per-category IoU distributions are shown in Figure 4, and the overall IoU values are considerably lower than in the experiment with COCO. For instance, the median IoU values for class `bottle` is below 0.8, and around 15% of the samples for this category present an IoU lower than 0.5. This means that an *ideal* object detector considering SHBB annotations would have an AP_{50} upper bound of approximately 85% for this category when validated with AHBB annotations.

Similarly to the experiment with COCO, Figure 5 shows a category-agnostic scatter plot of the IoU vs. the smallest SHBB dimension for all annotations in VOC. Compared to Figure 3, the IoU values for similar minimum dimensions are more spread and with smaller values. This behavior is actually expected, since the AHBB and SHBB annotation discrepancies go beyond subpixel differences in VOC. The AHBBs are not necessarily "bounding" representations of the objects compared to the segmentation mask, as illustrated in Fig. 1b. AHBBs for objects with partial occlusions are also not consistent: in some cases, the AHBB covers only the central portion of the object, but in others, it does encompass even regions of the object that are not visible and were guessed by the annotator, as illustrated in Fig. 1a. Nevertheless, we can still find an overall monotonic relationship between the IoU and the smallest dimension of the SHBB in Figure 5: the Spearman coefficient was 0.79, with a numerically null p-value.

For a better evaluation of the annotation discrepancies, we consider that SHBBs are the ideal annotations and compute the coordinate offsets (top-left and bottom-right) between the AHBB and the SHBB. Although these offsets might reach hundreds of pixels for some images, approximately 95% of them lie on the square region $[-16, 16]^2$,

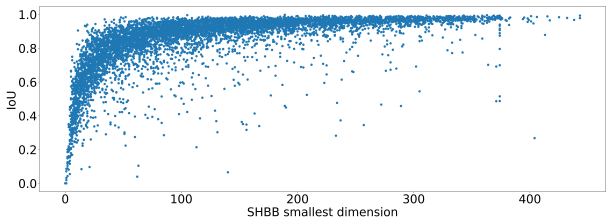


Figure 5. Relationship between IoU and smallest SHBB dimension for VOC.

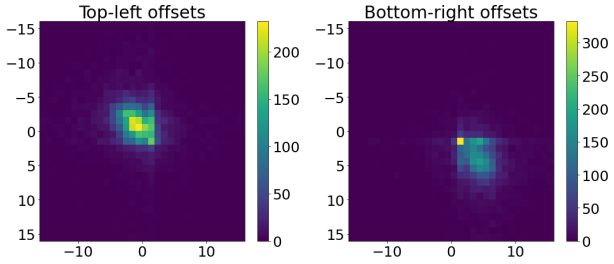


Figure 6. Offset between AHBBs and SHBBs for VOC.

shown in Fig. 6. If the AHBBs were indeed bounding representations of SHBBs, the top-left offsets would be all non-positive, and the bottom-right would be non-negative, which is not true in Fig. 6. In fact, 73.46% of the offsets satisfy $\Delta x_{tl} \leq 0$ and $\Delta y_{tl} \leq 0$, whereas 88.28% satisfy $\Delta x_{br} \geq 0$ and $\Delta y_{br} \geq 0$, with a total of 66.09% satisfying all constraints for a bounding representation.

Since the discrepancies in VOC are larger than the subpixel differences in COCO, we also expect a stronger AP_T reduction when training object detectors using AHBBs and evaluating using SHBBs (note that human-labeled HBBs are provided for VOC, unlike COCO). Although only AP_{50} is suggested for VOC [4], we also evaluated more restrictive thresholds to evaluate the degradation. The results shown in Table 2 indicate that even using $T = 0.5$ produces a 4% AP_{50} reduction w.r.t. an ideal detector, and the degradation is over 22% when considering the COCO AP metric. We also evaluated two popular HBB object detectors, namely EfficientDet D0 [22] and SSD300 [13] with ResNet50 backbone, both trained using the AHBBs of the PASCAL VOC 2007 train set and validated with AHBBs and SHBBs of the segmentation subset, for which we can compute the SHBBs. In the least restrictive scenario (AP_{50}), there was an accuracy drop of approximately 2% for EfficientDet and SSD300, respectively. The accuracy drop becomes even larger for tighter IoU thresholds, reaching 4.5% for EfficienDet and 6.3% for SSD300 when AP_{75} is considered. When we consider $AP_{50:95}$, the accuracy drops for the two detectors are 7.6% and 9.3%, respectively. It is also important to mention that recent approaches that propose novel localization loss functions for HBB object detectors report relatively small AP_T gains for VOC. For example, Distance-IoU [31] reports an AP_{75} of 56.34% for their loss vs. 54.74% using the IoU loss in the Pascal VOC 2007 test set, which means a 1.6% improvement. Such gain is considerably smaller than the 6.3% gap shown in our experiments by just changing the GT annotation. Results with VOC confirm that small objects are more susceptible to errors, and disparities between the SHBB and AHBB can harm or benefit the performance of detectors.

4. Evaluating dataset self-consistency for OBBs

It is well known that HBBs only provide a coarse approximation of the object shape (i.e., its segmentation mask), particularly for irregular or articulated objects [11]. Even roughly rectangular shapes might not be well represented by HBBs when rotations are considered. For example, a long and thin object with a 45° rotation will be represented by a roughly square HBB, and rotating the same object by 90° yields the exact same HBB.

Oriented Bounding Boxes (OBBs) are becoming a popular alternative to HBBs, but their application is still limited mostly to niche applications such as text spotting (ICDAR 2015 [9] and MLT 2017 [17] datasets) and object detection in aerial satellite images (HRSC 2016 [14] and DOTA v1 [27] datasets). The visual OBB definition for objects that present a naturally elongated format such as words/sentences or ships is straightforward, but not so easy for irregular squared/circular objects. On the other hand, OBBs can be extracted automatically from segmentation masks using algorithms that fit a minimum-area bounding rectangle to the shape [5]. Here, we focus our analysis on the DOTA 1.0 dataset, which contains OBB annotations for several objects in 15 different categories, and the segmentation masks provided in iSAID [26] based on the same RGB images from DOTA v1 (called simply DOTA from this point on for simplicity). We compare the minimum-area OBBs extracted from the segmentation masks (called SOBBs) with the manually annotated OBBs (called AOBBs) using a similar strategy as the experiment with VOC (iSAID typically provides more segmentation mask annotations than AOBBs in DOTA), yielding a total of 98k pairs of AOBBs/SOBBs.

Fig. 7 shows the per-category IoU distributions comparing AOBBs with the corresponding SOBBs for DOTA, and Fig. 8 shows a few visual examples (since high-resolution images are used in DOTA, we show here image crops illustrating only the object under consideration). We note that categories with an inherently rectangular shape, such as `tennis-court` (Fig. 8a), `ground-track-field` (Fig. 8b), `basketball-court`, and `soccer-ball-field` present large IoU values (median above 0.85). The IoU dis-

Detector	AP_{50}	AP_{55}	AP_{60}	AP_{65}	AP_{70}	AP_{75}	AP_{80}	AP_{85}	AP_{90}	AP_{95}	$AP_{50:95}$
Ideal	100.0 / 96.05	100.0 / 94.85	100.0 / 92.96	100.0 / 90.76	100.0 / 87.74	100.0 / 84.03	100.0 / 78.54	100.0 / 69.90	100.0 / 55.66	100.0 / 27.78	100.0 / 77.83
Efficientdet	67.28 / 65.49	64.78 / 62.75	62.32 / 60.20	58.78 / 55.80	53.82 / 49.82	48.47 / 43.98	40.07 / 35.26	28.77 / 22.59	13.23 / 9.16	1.86 / 1.15	43.83 / 36.22
SSD300	63.02 / 61.52	61.23 / 59.16	59.49 / 56.87	56.70 / 53.42	53.29 / 48.25	48.12 / 41.87	38.99 / 32.02	26.38 / 19.52	9.66 / 5.48	1.12 / 0.44	41.80 / 32.53

Table 2. AP_T values (%) for a different object detectors in VOC trained with AHBBs and evaluated with AHBBs/SHBBs

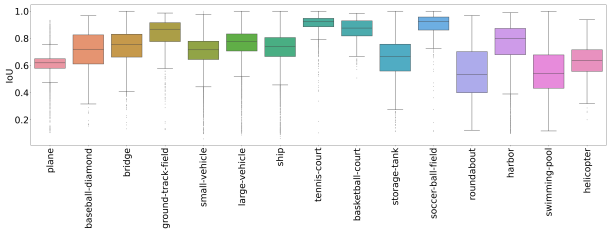


Figure 7. Iou between AOBBs and OBBs generated from segmentation masks (SOBBs) for DOTA/iSAID.

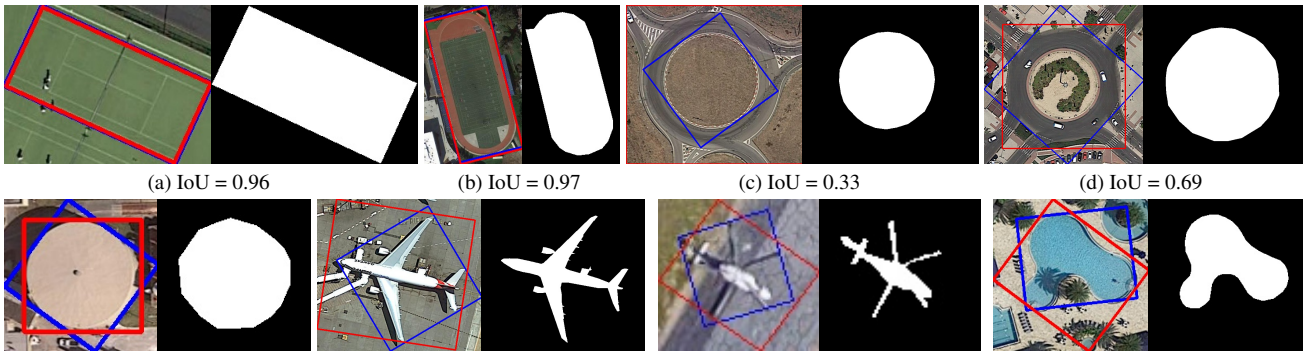
tribution for `roundabout` presented the smallest median IoU value (0.533), and the main cause was inconsistency in the segmentation annotation: in some cases, only the inner part of the roundabout was marked (Fig. 8c), whereas in the pavement around it was also included in the mask (Fig. 8d). Furthermore, perfectly circular masks generate ambiguous minimum-area bounding rectangles: any bounding square with arbitrary rotation presents the same area, and changing a single pixel in the mask can provide an artificially dominant orientation for the SOBB. The AOBBs, on the other hand, are drawn mostly as aligned OBBs (see Figs. 8c and 8d), which helps explaining lower IoU values. The circular issue also arises for `storage-tank`, as shown in Fig. 8e. A related behavior appears for categories `plane` and `helicopter`, which leads to a roughly square SOBB as shown in Figs. 8f and 8g, respectively. The orientation of the SOBB is rather arbitrary, and it depends on the shape of the aircraft. However, the AOBB presents a *semantic* orientation related to the main axis of the airplane or helicopter. At first glance, the relatively small IoU values for `swimming-pool` was a surprise, since we might think of rectangular shapes. However, there are many irregularly-shaped pools in the dataset for which the orientation is rather arbitrary, as shown in Fig. 8h.

Unlike the experiments with HBBs, we do not expect a clear monotonic relationship between the IoU and the smallest OBB dimension. As mentioned before, the IoU discrepancies between SOBBs and AOBBs are caused by other factors as well: human-centered biased when independently annotating segmentation masks or OBBs, and the ambiguity when generating OBBs from the segmentation masks, particularly for irregular shapes without a clear orientation, which is highly related to the object category). Nevertheless, we show a per-category scatter plot of IOU vs. smallest OBB dimension in Fig. 9, and note that some categories, such as `plane`, yield low IoU values even in larger OBBs. We can also note several samples from different categories that present low IoU values and a small SOBB.

Finally, we evaluate the impact of annotation discrepancies for OBB object detection based on AP_T metrics for ideal and real object detectors. Table 3 shows that the AP degradation for DOTA/iSAID is even more evident than VOC and COCO for HBBs. For an ideal detector, there was an almost 11% AP_{50} accuracy drop, and the AP_T values decay rapidly as the IoU threshold gets more restrictive, reaching 48.9% for AP_{75} and a mere 5.6% for AP_{95}, with an $AP_{50:95}$ of 51.6%. We also explored two SOTA OBB object detectors: OBB-adapted RetinaNet [28] and R^3 Det [29] with ResNet50 (R-50) backbone, both trained with the AOBBs in DOTA validated with both AOBBs and SOBBs. In the least restrictive scenario (AP_{50}), there was an accuracy drop of approximately 12% and 9% for RetinaNet and R^3det, respectively. The accuracy drop becomes even larger for tighter IoU thresholds, reaching 31% for RetinaNet and 27% for R^3det when the AP_{70} is considered. We can also note that the best detector for a given AP_T value varies depending on the chosen evaluation format. For example, RetinaNet performs better than R^3det in AP_{60} using AHBBs, but the opposite happens when considering SOBB annotations for the same AP level. In a nutshell, the experiments with OBBs indicate that discrepancies between direct OBB annotations and OBBs induced by segmentation masks are deeper than HBB annotations, affecting also objects with a roughly circular shape.

5. Conclusions

This paper presented a critical analysis of popular datasets for HBB and OBB object detection, namely COCO, VOC, and DOTA/iSAID, aiming to check the consistency of bounding box annotations and segmentation masks and how discrepancies affect the IoU and AP metrics. COCO does not present a set of human-annotated HBBs (AHBBs), and they are directly derived from the polygons that represent the segmentation mask (SHBBs). To emulate annotation errors, we simply rounded off the floating-point coordinates of the SHBBs and showed that even sub-pixel discrepancies between AHBBs and SHBBs can lead to strong IoU degradations, particularly for small objects. VOC presents independent sets of AHBBs and segmentation masks, from which we extracted the bounding boxes SHBBs. Our experiments indicate that AHBB and SHBB discrepancies might be due to minor annota-



(a) IoU = 0.96 (b) IoU = 0.97 (c) IoU = 0.33 (d) IoU = 0.69

(e) IoU = 0.71 (f) IoU = 0.57 (g) IoU = 0.61 (h) IoU = 0.61

Figure 8. Examples of OBB annotations (AOBBs) in DOTA (red) and the minimum enclosing rectangle related to the corresponding segmentation mask (SOBB, in blue).

Detector	AP_{50}	AP_{55}	AP_{60}	AP_{65}	AP_{70}	AP_{75}	AP_{80}	AP_{85}	AP_{90}	AP_{95}	$AP_{50:95}$
Ideal	100.0 / 89.30	100.0 / 84.53	100.0 / 77.84	100.0 / 68.31	100.0 / 58.78	100.0 / 48.87	100.0 / 38.23	100.0 / 27.53	100.0 / 16.81	100.0 / 5.56	100.0 / 51.58
RetinaNet	82.20 / 70.62	81.58 / 66.04	80.57 / 58.13	78.54 / 51.00	75.12 / 44.37	66.99 / 36.25	55.00 / 26.65	37.11 / 16.18	16.40 / 7.74	2.78 / 0.95	57.62 / 37.79
R^3det	79.90 / 70.77	79.05 / 65.68	77.47 / 58.32	74.89 / 51.19	69.27 / 42.81	59.90 / 34.16	45.56 / 24.41	27.82 / 14.84	9.96 / 5.82	0.58 / 0.30	52.44 / 36.83

Table 3. AP_T values (%) for a different object detectors in DOTA/iSAID trained with AOBBs and evaluated with AOBBs/SOBBs

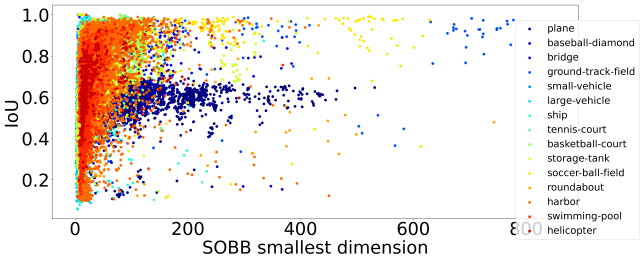


Figure 9. Scatter plot showing the IoU between AOBB/SOBB pairs vs. the smallest SOBB dimension for DOTA/iSAID.

tion inaccuracies but also to human-centric views on exactly what is considered the object of interest. As expected, AHBB/SHBB discrepancies were larger than COCO, and so was the IoU degradation. Finally, we considered the AOBB annotations in DOTA, and generated a set of corresponding segmentation-induced OBBs (SOBBs) from the related dataset iSAID. We observed that AOBB/SOBB discrepancies were even higher than HBB datasets, in part due to human-centric views (as in VOC) but also due to the orientation ambiguity in roughly circular or irregularly shaped objects that lead to approximately square boxes with no clear orientation.

We also performed experiments by using ideal or real HBB and OBB object detectors that are trained using either annotated boxes or segmentation-induced boxes, and evaluated using both representations. We observed non-neglectable degradation of the AP metrics in the cross-representation experiments (i.e., training with one representation and evaluating with the other), particularly for tighter IoU thresholds and smaller objects. Although the AP degradation arises even in sub-pixel annotation discrepancies, as shown in the experiments with COCO, it is more evident when the discrepancies are larger (experiments with VOC) and even stronger when OBB representations are used (experiments with DOTA/iSAID).

Our results indicate that the blind use of IoU (which impacts the widely adopted AP metrics) for comparing bounding boxes is dangerous, and results might be strongly affected by annotation errors or human-centric bias – particularly for small objects. As well-noted in [20], available annotations are only *approximations* of the actual GT: requiring tight adherence to these approximations measured by the IoU might not necessarily mean tight adherence to the *actual* and unfortunately unknown GT. Given that annotation errors affect boxes differently according to their sizes, one alternative for validating object detectors would be to adjust the IoU acceptance threshold based on the individual BB dimensions: more flexible thresholds could be used for smaller boxes, and more restrictive ones for larger boxes. We also suggest that new datasets provide annotations of multiple humans for the same images, allowing a more reliable estimate of inter-annotator discrepancies and their effect on the IoU of different categories and object sizes.

6. Acknowledgments

We thank the financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001, Brazil. We also thank the Google Cloud Research Credits Program.

References

[1] Jia Deng. A large-scale hierarchical image database. *Proc. of IEEE Computer Vision and Pattern Recognition, 2009*, 2009. 2

[2] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 5

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 2, 3, 6

[5] Herbert Freeman and Ruth Shapira. Determining the minimum-area encasing rectangle for an arbitrary closed curve. *Communications of the ACM*, 18(7):409–413, 1975. 6

[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. 2

[7] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1031–1040, 2020. 3

[8] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report. 1

[9] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160, 2015. 1, 6

[10] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 3, 4, 6

[12] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020. 1

[13] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 6

[14] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *the 6th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, pages 324–331. INSTICC, SciTePress, 2017. 1, 6

[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. International Conference on Computer Vision*, volume 2, pages 416–423, July 2001. 2

[16] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939, 2016. 2

[17] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, Wafa Khlif, Muhammad Muzzamil Luqman, Jean-Christophe Burie, Cheng-lin Liu, and Jean-Marc Ogier. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1454–1459, 2017. 1, 6

[18] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. 2

[19] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 2

[20] Hamid Rezatofighi, Tran Thien Dat Nguyen, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustworthy are performance evaluations for basic vision tasks? *arXiv preprint arXiv:2008.03533*, 2020. 2, 3, 8

[21] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 3

[22] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 4, 6

[23] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020. 2

[24] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021. 4

[25] Zining Wang, Di Feng, Yiyang Zhou, Lars Rosenbaum, Fabian Timm, Klaus Dietmayer, Masayoshi Tomizuka, and

Wei Zhan. Inferring spatial uncertainty in object detection. In *IEEE/RSJ IROS*, pages 5792–5799. IEEE, 2020. 3

[26] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAID: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 6

[27] 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 *CVPR*, June 2018. 1, 6

[28] Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, and Junchi Yan. Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15819–15829, June 2021. 7

[29] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *AAAI*, 2021. 7

[30] Xue Yang, Junchi Yan, Ming Qi, Wentao Wang, Zhang Xiaopeng, and Tian Qi. Rethinking rotated object detection with gaussian wasserstein distance loss. In *ICML*, 2021. 1

[31] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12993–13000, 2020. 6