Oriented Cell Dataset: A Dataset and Benchmark for Oriented Cell Detection and Applications

Lucas Kirsten[1]*, Angelo Angonezi[2], Jose Marques[1], Fernanda Oliveira[2], Juliano Faccioni[2], Camila Cassel[2], Débora de Sousa[2], Samlai Vedovatto[2], Guido Lenz[2], Claudio Jung[1]
Federal University of Rio Grande do Sul, Brazil – [1]Institute of Informatics, [2]Institute of Biosciences

Abstract

This work presents a new public dataset for cell detection in bright-field microscopy images annotated with Oriented Bounding Boxes (OBBs), named Oriented Cell Dataset (OCD). Our dataset also contains a subset of images with five independent expert annotations, which allows inter-annotation analysis to determine a suitable IoU acceptance threshold for evaluating cell detectors. We show that OBBs and a derived representation, Oriented Ellipses (OEs), provide a more accurate shape representation than standard Horizontal Bounding Boxes (HBBs) with a slight overhead of one extra click in the annotation process. We benchmarked OCD using 14 state-of-the-art oriented object detectors, and explored two main problems in cancer biology: cell confluence and polarity determination. Our code and dataset are available at https://github.com/LucasKirsten/Deep-Cell-Tracking-EBB.

1. Introduction

Detection and tracking of living cells in microscopy images is a crucial task required in many medical and research applications, such as cell growth, migration, invasion, and morphological changes [3, 8, 21, 30]. Bright-field microscopy has several advantages, such as not requiring any fluorescent tagging of the cell, reduced photo-toxicity, and much more affordable microscopes [29]. The drawback of bright-field images is the difficulty of automating the analysis due to the lower contrast to the background compared to fluorescence microscopy, and images might contain artifacts similar to the cells. Multiple approaches have been proposed to detect or segment cells in bright-field microscopy images [1], but the diversity of microscopes, cell lines, and level of magnification hinders the possibility of applying the same pipeline to different experimental setups.

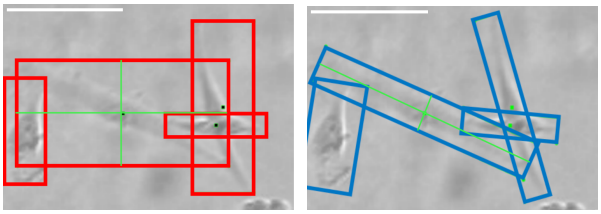
Besides identifying cells, measuring their size and shape is relevant for identifying phenotypes [7] and migration patterns [30]. Horizontal Bounding Boxes (HBBs) are the *de facto* choice for generic object detection [46] and cell detection [1]. However, they are not suited to obtain the actual shape and size of the cells, particularly for elongated and orientated cells (see Figure 1a). Obtaining the full cell masks through segmentation approaches is an alternative, but the annotation process is tedious. Recent foundation models such as SAM [16] can leverage automatic or semi-automatic segmentation. Still, the definition of the masks is ill-defined for overlapping cells [12] (see Figure 1) or low-contrast images (see Figures 2c and 2d). In this work, we advocate that Oriented Bounding Boxes (OBBs) are adequate representations for cellular imagery applications, with a good compromise between object representation and annotation cost. We introduce the **"Oriented Cell Dataset" (OCD)**, which provides annotated bright-field cell images as OBBs, and perform a thorough benchmarking with state-of-the-art oriented object detectors. We also use the results of these detectors to analyze two important biological applications: cell confluence and polarity estimation. Our main contributions are: i) we introduce OCD, the first dataset containing bright-field cell images annotated as OBBs; ii) using the popular Cell Tracking Challenge (CTC) [23], we show that oriented representations such as oriented bounding boxes (OBBs) and oriented ellipses (OEs) present considerable more overlap with the segmentation masks than traditional horizontal bounding boxes (HBBs); iii) we perform an inter-annotator assessment (IAA) evaluation of OBB human annotations to estimate the human variability, and propose a suitable IoU threshold for evaluating automatic OBB cell detectors; iv) we show that oriented object detectors can be employed to automate biological tasks of cell detection, such as confluence and polarity estimation.

2. Related work

2.1. Cell Microscopy Datasets

Public datasets are crucial in advancing cell detection methodologies, offering diverse and comprehensive benchmarks for evaluating and training deep learning mod-

*Corresponding author: lnkirsten@inf.ufrgs.br



(a) (b)

Figure 1. Comparison of (a) HBB and (b) OBB annotations for glioblastoma cells. Green lines show bounding box axes, which are used to calculate the cell area and polarity. For the middle cell, the area is ∼2.6 times larger whereas the polarity is ∼3 times smaller when HBB and OBB of the same cell are compared. Scale bar (top-left white) equals 30 μm.

els. The Cell Tracking Challenge (CTC) dataset [22, 23] is a standard literature benchmark, providing 48 time-lapse sequences split evenly between training and competition phases. These sequences include real and computer-simulated videos, with various cell densities and noise levels. The dataset offers segmentation annotations (cell instance masks) and tracking annotations (cell markers interlinked between frames), categorized into gold standard (majority opinion from experts) and silver standard (computer-generated annotations). This distinction is essential as the silver-truth significantly enhances cell instance coverage, reducing the reliance on labor-intensive manual annotations, which only cover 17.8% of cell instances in the gold standard. The dataset by Ker et al. [15] comprises 48 phase-contrast time-lapse microscopy sequences of Mouse C2C12 cells under various treatments. The annotations are manually performed, marking cell centroids approximately every 1 to 8 frames and interpolating the cell position between these frames. Despite the annotations requiring only cell markers, the process is extremely time-consuming, taking up to 1.5 years to complete, underscoring the challenges associated with manual annotation. More recently, the CMTC dataset [2] provides 86 videos from 14 cell lines with 152,584 frames. These videos are annotated by drawing HBBs around cells and using linear interpolation to reduce manual effort, facilitating more efficient annotation.

These datasets highlight the critical balance between annotation quality and labor, hinting at developing more efficient and effective cell annotation schemes. HBB annotations are fast to produce, but capture only a rough estimate of the cell shape. Full segmentation masks provide a complete shape description, but are costly to annotate – pseudo-masks can be used, but they might contain errors. In this work, we advocate that OBB annotations provide a good cost-benefit regarding annotation time and shape representation for cell imagery. To the best of our knowledge, we present the *first cell dataset containing manually annotated OBBs*. We also perform a thorough investigation on state-of-the-art (SOTA) oriented object detectors to automatize cell research.

2.2. Oriented Object Detection

As noted in [1], a popular approach for cell detection in microscopy images is to use specialized versions of algorithms for general-purpose object detection/segmentation, such as Faster-RCNN [31]. Both HBB and OBB detectors share the same main concepts in terms of network architecture (i.e., both can be achieved using two- or one-stage methods), and the main differences relate to the regression head: an additional parameter related to angular information must be produced by the detector. Next, we revise some SOTA generic-purpose OBB detectors that can be re-trained for oriented cell detection.

Two-Stage Methods: In two-stage methods, the first stage creates OBB proposals, and the second predicts the class-related confidence for each proposal and refines its shape. With horizontal proposals often leading to mismatches between Regions of Interest (RoI) and objects, early two-stage oriented detectors use rotated anchors to generate Rotated RoIs (RRoIs), increasing the computational complexity and memory footprint of models. For instance, Han et al. [11] explores a backbone capable of extracting rotation-equivariant features and a novel Rotated RoI Align for extraction of rotation-invariant features. Xie et al [38] addressed the inherent high computational cost of [5] with Oriented R-CNN, capable of generating high-quality oriented proposals in a nearly cost-free manner.

One-Stage Methods: One-stage methods aim to simultaneously regress an OBB related to an object and predict its class label, increasing computational efficiency. For instance, Lin et al [19] proposed RetinaNet, a simple yet effective single-stage anchor-based architecture with a novel classification loss function, which was adapted for OBB detection in [39]. Yang et al [40] used a feature refinement module, which re-encodes refined bounding boxes to their corresponding feature points through pixel-wise feature interpolation. Han et al [10] later propose S2ANet, which gets rid of heuristically defined anchors by refining horizontal anchors into high-quality rotated ones, doing so in a fully convolutional way. Other approaches, called anchor-free, do not explore the idea of anchors. For example, FCOS [34], regresses bounding boxes from a center point and filters predictions using a predicted centerness value, while Oriented RepPoints [35] uses adaptive points coupled with an effective adaptive points assessment and assignment scheme for measuring the quality of generated points.

Regression loss functions for OBBs: Another critical distinction between HBB and OBB detection is the choice for the regression loss. Although parameter- or corner-wise terms using different norms (e.g., ℓ_1, ℓ_2, Huber) can be used, a recent trend is to explore "holistic" terms that di-

rectly compare the OBBs [9, 44]. Nevertheless, new works explore the relations between OBBs with 2D Gaussian functions and leverage statistical distance measures as the regression loss, such as the Gaussian Wasserstein Distance (GWD) [41], the Kullback-Leibler Divergence (KLD) [42], and the Probabilistic IoU (ProbIoU) [25].

Although these methods have been broadly used for detecting oriented objects in aerial/satellite images [36, 43], we are unaware of their application for oriented cell detection. In this work, *we provide a thorough benchmark of SOTA OBB detectors in the proposed OCD dataset and explore the results to estimate cell confluence and polarity*, which are broadly used in biological applications [8,21,33]. These applications require area and orientation information, and HBB or point-wise annotations are not suitable.

3. The Oriented Cell Dataset (OCD)

We created a new public cell dataset[1], called *Oriented Cell Dataset (OCD)*[2], containing OBB annotations for bright-field microscopy images. A total of 160 images were acquired using different microscopes and cell lines, which resulted in visually distinct images as shown in Figure 2. These images were split into three sets:

• **Train**: 120 images (75% of total), used for training the object detectors, with 4,602 annotated OBBs;

• **Test**: 30 images (18.75% of total) used for evaluating the object detectors, with 992 annotated OBBs;

• **Annotator's Comparison**: 10 images (6.25% of total) used for the inter-annotator assessment (see Section 4.2), with an average of 326 OBBs per annotator.

Each image of the *Train* and *Test* splits was annotated by a single annotator, and they contain an equal distribution (stratified) regarding the used microscope, cell lines, and cultivation method. Meanwhile, five different human experts annotated each image of the *Annotator's Comparison* split. Although our dataset is relatively small compared to other available benchmarks [2, 15, 22, 23], we highlight that our data is composed of only *real cancerous human cells* and is *entirely manually labeled and reviewed by human experts*. A full description of the dataset acquisition and splits is provided in Table 1.

The images were manually annotated using the roLabelimg tool[3], which allowed the researchers to delimit the OBBs (composed of x-center, y-center, height, width, and angle) for each cell. Furthermore, the cells were classified either as "elongated" or "round" (corresponding to regular or mitotic cells, respectively), as illustrated in Figure 3. In total, over 6,700 OBBs were annotated, from which 89.4% were classified as "elongated" and 10.6% as "round" cells. The confluence (i.e., the area occupied by cells in a given image I) was computed as

$$\mathcal{C} = \frac{\sum_{i=1}^{N_I} \#\text{OBB}_i}{\#\text{I}}, \tag{1}$$

where $\#\text{OBB}_i$ is the OBB area of cell i, N_I is the number of cells in image I, and $\#$I is the image area. We obtained a mean confluence among the images of $\mathcal{C}$ ≈10%, ranging from nearly 0% to 40%.

Considering the *Train* and *Test* splits, which were used to effectively train and evaluate the models, 42.2% of the OBBs are small, 50.5% are medium, and only 7.3% are large based on the area thresholds (32^2 and 96^2, respec-

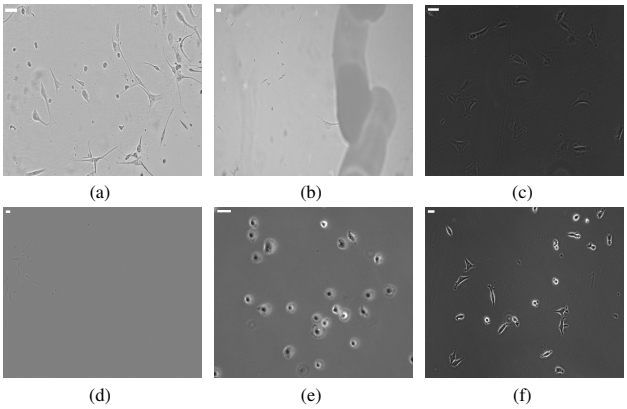


Figure 2. Example of different images from the OCD. (a) and (b) are A172 cells captured in CytoSMART microscope; (c) and (f) are MRC5 and MCF7 cell lines, respectively, captured in Zeiss Axiovert 200 microscope; (d) are U251 cells captured in IncuCyte; and (e) are SCC25 cells captured in a Zeiss AxioCam microscope. Scale bar (top-left white) equals 30 μm.

Table 1. OCD description. Cell lines A172 and U251: human glioblastoma; MCF7: human breast cancer; MRC5: human lung fibroblast; SCC25: human squamous cell carcinoma. Cultivation condition CTR: cells in the control group - without the addition of chemotherapy; TMZ: cells treated with 50 μM temozolomide in some cultivation step - 3h treatment in most experiments.

Split	Microscope	Magnification	Cell line	Cultivation Condition	# imgs.	Images Resolution
Train	CytoSMART	10x/20x	A172	CTR	52	1280×720
				TMZ	25	
			U251	CTR	4	
	Zeiss Axiovert 200	10x	MCF7	CTR	19	1388×1040
			MRC5	CTR	20	
Test	CytoSMART	10x/20x	A172	CTR	4	1280×720
				TMZ	5	
			U251	CTR	1	
	Zeiss Axiovert 200	10x	MCF7	CTR	10	1388×1040
			MRC5	CTR	10	
Annotators' Comparison	CytoSMART	10x/20x	A172	CTR	4	1280×720
				TMZ	1	
	IncuCyte	10x	U251	TMZ	1	1408×1040
	Zeiss AxioCam	20x	SCC25	CTR	2	600×512
	Zeiss Axiovert 200	10x	MCF7	CTR	1	1388×1040
			MRC5	CTR	1	

[1]Generated at Labsinal: Cell Signaling and Plasticity Laboratory – UFRGS (https://www.ufrgs.br/labsinal/)

[2]Available at: https://ieee-dataport.org/documents/oriented-cell-dataset-ocd.

[3]https://github.com/cgvict/roLabelImg

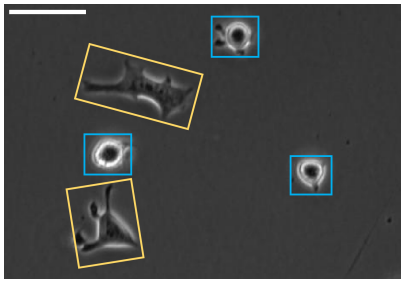


Figure 3. Illustration of "elongated" (yellow) and "round" (blue) cells in the OCD. Scale bar (top-left white) equals 30 μm.

tively) of the COCO evaluation protocol (see Figure 4a). As noted in [4], object detectors produce smaller evaluation metrics for small objects, which makes OCD a challenging dataset regarding object size. The histogram of OBB aspect ratios (largest and smallest dimension) is shown in Figure 4b. Almost 60% of the annotated OBBs present an aspect ratio larger than two. The per-image cell count distribution, shown in Figure 4c, indicates a median value of 24.5, with only 10 images containing more that 100 cells.

4. Experiments and Applications

In our experimental setup, we first evaluate if OBBs are suitable approximations for the cell shape. Then, we perform an inter-annotator assessment (IAA) experiment to evaluate the degree of agreement among the human annotators and discuss suitable acceptance thresholds for computing the associations between predicted detections and ground-truth (GT) annotations. Using the *Train* split of our dataset, we train object detectors to asses if their variability is similar to what is expected among human annotators. Finally, we illustrate the usefulness of the OCD in two biological applications: cell confluence and polarity estimation.

4.1. Are OBBs adequate representations?

To evaluate if OBBs are good approximations for the cell shapes, we evaluated seven public datasets from the Cell Tracking Challenge (CTC) [23][4], which contain annotated segmentation masks provided as *silver* and *gold* standards. The silver standard annotations refer to computer-origin reference annotations, while the gold standard refers to human-origin ones. Since only 17.8% cells are annotated in the gold standard, most works use only the silver standard annotations, which emphasizes the difficulty of manually annotating segmentation masks on microscopic images.

We considered all segmentation masks (silver and gold standards) as the reference (GT) cell shape representation and fitted both HBB and OBB representations as approximations of the mask based on provided OpenCV[5] implementations. We also considered the natural extensions of these approximations for fitting elliptical-like objects: converting an HBB to an Axis-Aligned Ellipse (AAE) with major and minor axis as the major and minor HBB lengths; and the OBB to an Oriented Ellipse (OE) with major and minor axis as the major and minor OBB lengths, respectively. Both center points and angle of rotation (in the OBB case) are kept the same in the AAE and OE representations. We used the Intersection over Union (IoU) between the reference shape representation (i.e., the segmentation mask) and the four approximate representations (HBB, OBB, AAE, or OE) as the quality metric. Furthermore, we computed the per-annotation average of the minimal number of points required to annotate the cells of each dataset using the segmentation masks to measure the annotation efforts.

[4] http://celltrackingchallenge.net
[5] https://opencv.org/

4.2. Inter-annotator assessment

In generic-purpose object detection, the presence, category, and HBB/OBB annotation for each instance are typically well-defined. On the other hand, annotating biomedical data is far more challenging, and experts might disagree with some of the data. For example, Figure 5 shows the annotations provided by different humans for the same image, with arrows indicating particularly challenging situations, such as blurry cells/background and clutter. We note that annotators might disagree both on the presence/absence of a cell and its exact shape. As such, the GT annotations are expected to comprise a "human variability" factor [24], which can also impact the objective metrics used to compare object detection approaches [27]. In this section, we propose to evaluate the level of agreement among human annotators regarding cell detection and classification (a.k.a. Inter-Annotator Assessment, or IAA).

We used a smaller subset of ten images that were independently annotated by five researchers (having multiple annotators for the full dataset would be unfeasible), composing the *Annotators' comparison* data split (recall Section 3). Inspired by [26], we explore the Krippendorff's Alpha ($K-\alpha$) metric for IAA. As in [26], we tackle the problem from the semantic segmentation perspective, where each image pixel presents a single label within a set of X values inherited from the OBB annotations. We propose two different categorizations in our analysis: i) a *class-aware* version, where each pixel is labeled as background, elongated, or round cell ($X = 3$); and ii) a *class-agnostic* version, where each pixel is labeled as either background or cell ($X = 2$). For more details regarding $K-\alpha$, we refer the reader to [26].

The Krippendorff's Alpha score provides an overall agreement metric among the annotators, but the score might be dominated by background pixels. Furthermore, it is not able to distinguish overlapping cells. On the other hand, the agreement of the OBBs annotated by two experts for the

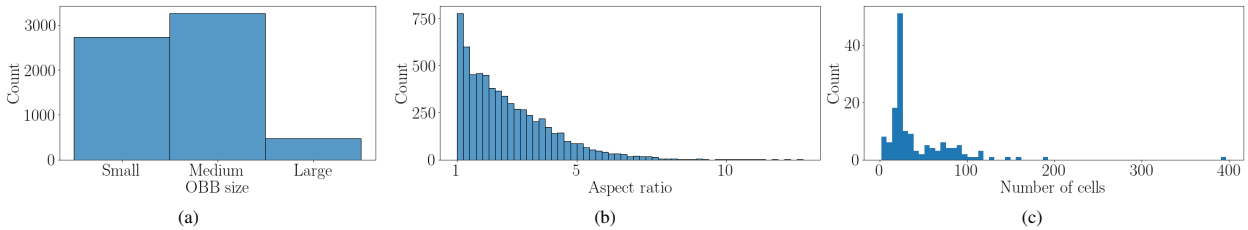


Figure 4. (a) Histogram of (a) OBB sizes, (b) OBB aspect ratios and (c) per-image number of cells in the *Train* and *Test* splits.

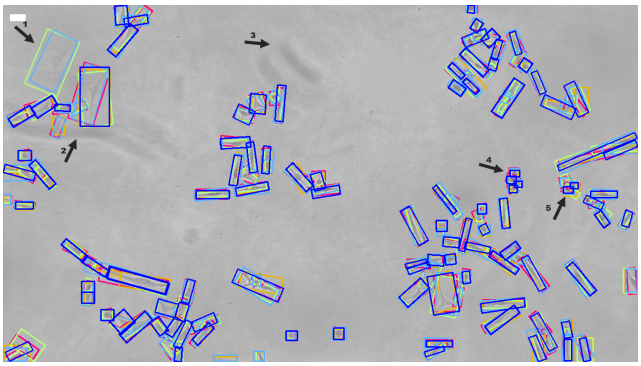


Figure 5. *Annotator's Comparison* dataset image example. OBBs are colored according to each different human annotator. Arrows indicate challenging situations that might hinder the models' performance, such as (1) cells out of focus, (2,3) blurry background, (4,5) and aggregated cells. Scale bar (top-left white) equals 30 μm.

same cell is crucial to evaluate the precision, recall, and AP metrics, which are the standard metrics for generic-purpose object detection [20] and can be strongly affected by choice of the IoU threshold [27]. To evaluate the individual effect of cell shape annotation, we performed a second experiment with the *Annotators' comparison* data split.

More precisely, let us consider $\mathcal{A}_i$ and $\mathcal{A}_j$ the set of annotated OBBs related to annotators i and j, respectively, and let N_i and N_j denote the corresponding number of annotated cells for a given image. For each pair of annotators $i \neq j$, we consider $\mathcal{A}_i$ as the GT and $\mathcal{A}_j$ as the set of candidate detections. Since there is no confidence associated with the OBBs, we cannot compute the AP metric [20]; instead, we use the F1-Score to evaluate the detection set, varying the IoU threshold T to validate a candidate detection. Since our goal is only to evaluate the geometric consistency of the OBBs, we performed the analysis in a class-agnostic manner only (i.e., both elongated and circular cells are considered in the same category).

Since typically $N_i \neq N_j$, an analysis with the raw (unfiltered) annotations provides the joint effect of T, false positives (FPs), and false negatives (FNs) due to lack or excess of annotations. To isolate the individual effect of the threshold T, we have also performed an analysis with a filtered set of detections, obtained by applying the Hungarian Algorithm [17] between every pair of annotation sets $\mathcal{A}_i$ and $\mathcal{A}_j$ for all images using $1 - \text{IoU}$ as the association cost. The optimal association path matches $\min\{N_i, N_j\}$ pairs of OBBs, but it might match OBBs with very small (or no) intersections that probably relate to different cells. To avoid such mismatches, we eliminated pairs of matched OBBs with an IoU smaller than a threshold T_{min}, set empirically to 0.1.

4.3. OBB detectors for oriented cell detection

We have performed a thorough benchmark of the proposed OCD dataset using SOTA oriented object detectors. For the sake of reproducibility, we selected detectors that provide publicly available implementations using as a base source the MMRotate[6] toolkit [45], which allows several combinations of models and regression loss functions. More precisely, we used the baseline detectors RetinaNet [19], R^3Det [40], S2A-Net [10], Oriented RepPoints [35], ReDet [11], Oriented RCNN [38], FCOS [34]; and explored some of them combined with alternative regression loss functions such as GWD [41], KLD [42], and ProbIoU [25] based on code availability. All detectors were trained using the *Train* split of OCD, and we used the *Test* split to compute the evaluation metrics. We used the default hyper-parameters as designed for the DOTA [37] dataset to train all the models in the MMRotate[7].

We used the Average Precision (AP) [20] standard benchmark metric for object detection to evaluate the trained models. This metric relies on the IoU between annotated and predicted cells, which involves an acceptance threshold. In this work, we selected this acceptance threshold as 0.5 based on the IAA experiment, as described in Section 5.2. To evaluate the broader problem of "cell detection" without considering the category (round or elongated), we also performed a class-agnostic version of the AP.

[6] https://github.com/open-mmlab/mmrotate

[7] Refer to this repository for the settings: https://github.com/jhlmarques/OCDDataset.

4.4. Biological applications

Object detection is typically not a goal but an intermediary result from which relevant information can be extracted. This work focuses on two biological applications that can benefit from oriented object detection, as explained next. Unlike the AP, practical applications of object detectors require the definition of a detection threshold. For all detectors, we used a canonical detector score of 0.5, noting that this might not be the optimal value.

Cell confluence refers to the relative area occupied by cells in a given image. It is a crucial task in migration and wound-healing assays [8,21]. We run all the trained OBB detectors in the *Test* data split and compute the per-image estimated confluence using Eq. (1). Results are evaluated quantitatively against the GT values using the Mean Relative Error (MRE) across all images.

Cell polarity relates to an asymmetric and ordered distribution along an axis [14], which plays an important role for analyzing cell migration patterns [30]. Following [30], we use the major OBB axis as the cell orientation and quantify the polarity as the ratio between the major and minor axes. Using all trained OBB detectors, we estimated the polarity for each cell in a given image. Then, each image was characterized by the polarity histogram of all cells in the image, computed with a bin width of 0.5. For a quantitative comparison of the detectors, we used the χ^2 distance between the normalized histogram P from the detector and Q from the GT annotations, given by $\chi^2(P,Q) = \sum_i (P_i - Q_i)^2/Q_i$, and reported the average χ^2 values across all images for each detector.

5. Results

5.1. Comparison of cell representations

We present the results for representation comparison experiments described in Section 4.1 with the public CTC [23] datasets in Table 2. We can observe that the OBB representation provides consistently larger IoU values than HBBs, which is expected since HBBs might contain significant portions of the background. More precisely, the OBB representation provides 20.7% larger IoU values than HBBs on average, and 4.7% better than the AAEs. When using OEs, the improvement is even higher: 37.7% when compared to HBB, and 19.3% to AAE. This can be related to the fact that cells in several lineages usually present a more "round-like" shape, similar to a rotated ellipse.

Regarding the annotation efforts of segmentation masks, we can observe that OBB (and OE) representations require three points (clicks from the user on the screen to annotate), but polygonal representations related to full segmentation masks might scale this number from 27 up to 405 points (last column of Table 2), *which corresponds to 9 to 135 times more labor required from the human experts*. Hence, we advocate that cells can be annotated using OBBs, but the analysis can be performed using OEs to better overlap with the segmentation masks.

Table 2. Comparison of different representations for approximating the segmentation masks on the CTC [23] datasets. The first four columns contain the IoU, and the last column (SEG Pts.) refers to the average minimal number of points required to define one segmentation mask on the dataset images.

	IoU (in %)				SEG
Dataset name	**HBB**	**AAE**	**OBB**	**OE**	**Pts.**
DIC-C2DH-HeLa	65.11	74.92	71.25	83.79	216
Fluo-C2DL-Huh7	58.82	68.56	66.79	77.42	168
Fluo-C2DL-MSC	22.01	24.11	33.34	34.97	405
Fluo-N2DH-GOWT1	74.09	90.02	77.29	94.31	107
Fluo-N2DH-SIM+	66.46	80.87	72.89	88.75	81
Fluo-N2DL-HeLa	70.60	82.54	78.61	89.62	27
PhC-C2DL-PSC	44.52	49.22	64.47	70.01	27

5.2. Inter-annotator assessment

Following the experimental setup described in Section 4.2, we initially computed the $K-\alpha$ for the five human annotators. In the class-aware experiment, we obtained $K-\alpha = 0.760$; in the class-agnostic experiment (where all cell annotations were considered as the same category) the $K-\alpha$ increased to 0.794. As noted in [13], a $K-\alpha$ value above 0.67 is considered good for IAA in the industry and academia, and above 0.8 is very good. Hence, we conclude that the agreement among the human annotators is between good and very good for both class-aware and class-agnostic experiments, being particularly better in the class-agnostic version due to label discrepancies for some cells.

The second analysis described in Section 4.2 focuses on the impact of the IoU acceptance threshold over the quality metrics, consolidated through the F1-Score. We consider all pairwise combinations of annotators using both the raw (unfiltered) OBBs and the filtered version that considers only OBBs that were marked by both annotators. Figure 6 shows the plot of the mean F1-Score as a function of the IoU threshold T for the two scenarios (unfiltered and filtered), and the shaded regions indicate the variability when changing the pairs of compared annotators. As we can observe, the peak average F1-Score for the unfiltered experiments is around 0.8, which can be explained by FPs and FNs that are detected due to disagreements regarding cell, debris, or background. On the other hand, the results with the filtered annotations evaluate the individual effect of T for comparing the OBB of cells that were effectively marked by both experts but still might generate discrepancies in the shape and location. As we can see in Figure 6, the results are very consistent for lower IoU thresholds but rapidly decay when more restrictive thresholds are used. This result clearly indicates that using the widely adopted COCO metrics [20],

which considers a range of IoU values from 0.5 to 0.95, *is not realistic* for the cell detection problem using OBBs. In fact, the average IoU values considering all paired OBBs with the Hungarian Algorithm was 0.69, which is much lower than similar experiments using HBBs and generic-purpose datasets: 0.88 in [28] and 0.87 in [18].

To provide an IoU threshold that considers the inherent annotation variability, we computed the IoU value for which the relative cost of increment is no longer worth the corresponding performance benefit (i.e., the IoU value starts to be so restrictive that harms the F1-Score between pairs of human annotators) by finding the "knee-point" for the unfiltered and filtered data[8] [32], which both arise around a threshold of $T = 0.55$. Since this value was obtained based on expert annotations, we suggest using it as *an upper bound* for oriented cell detectors. Our findings are more aligned with the single threshold of 0.5 suggested in the Pascal VOC Challenge for object detection [6], which is used to evaluate the trained detectors next.

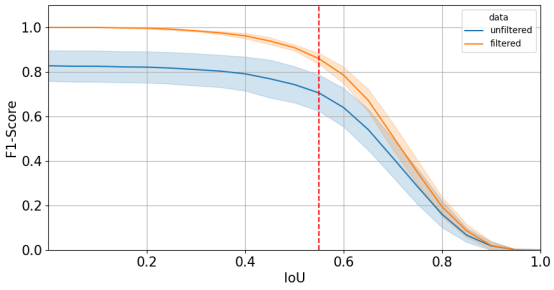


Figure 6. IAA evaluation based on F1-Score in different IoU thresholds. Dark lines refer to F1-Score mean values, shadow area indicates standard deviation, and the vertical dashed red line intersection with the curves refers to the curve's knee-points.

5.3. Cell detection results

Following the experimental setup described in Section 4.3, we trained 14 OBB detectors in OCD. The evaluation was performed using class-aware (CAw) and class-agnostic (CAg) AP values with an IoU acceptance threshold 0.5. These results are shown in the first two columns of Table 3. For some methods, the improvement from the CAw to CAg evaluation strategies was large (around 10% for Oriented RepPoints and ReDet), indicating that cell OBBs were correctly regressed but some category labels were swapped. Overall, Oriented RepPoints achieves the largest AP value in both CAw and CAg strategies. The second-best detectors for CAw and CAg evaluation were S2ANet and ReDet, respectively. We can also note that R^{3}Det benefited from the use of alternative regression loss functions (GWD, KLD, ProbIoU) in both CAw and CAg protocols. RetinaNet presented a similar behavior for the CAw protocol regarding alternative loss functions, but not in the class-agnostic experiments.

Examples of OBB detection with the best method – Oriented RepPoints – are shown in Figure 7. The image on the left illustrates a challenging high-confluence scenario, with several overlapping cells that are hard even for humans to identify, while the image on the right illustrates a sparser scenario. Although it is not easy to visually assess these results, we note that the detected cells match fairly well the annotations, particularly for the right image. We also note some false negatives related to round cells, indicated by arrows. However, recall that even the experts might mistake debris for cells and vice-versa.

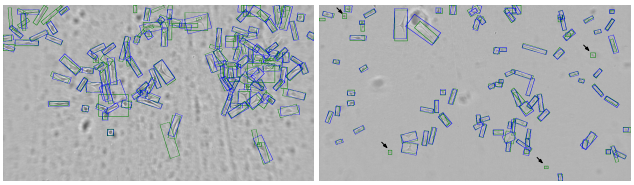


Figure 7. Examples of OBB detection with Oriented RepPoints (blue) and GT annotations (green).

Table 3. Results in the *Test* data split. Cell detection was measured using AP_{50}, in class-aware (CAw) and class-agnostic (CAg) regimes. Confluence evaluated with the mean relative error (MRE), and polarities using the χ^2. Best values are in **bold**, and second-best underlined.

	Object Detection		Confluence	Polarity
Method	**CAw** AP ↑	**CAg** AP ↑	**MRE** ↓	χ^2 ↓
FCOS	72.42	78.93	0.3243	12.2906
Oriented RCNN	74.11	79.14	0.1029	7.5440
Oriented RepPoints	**78.98**	**86.28**	0.1333	**5.5738**
R^{3}Det	68.64	75.80	0.1958	8.8894
R^{3}Det + GWD	69.76	77.95	0.1327	7.8373
R^{3}Det + KLD	72.02	78.37	0.1337	8.9585
R^{3}Det + ProbIoU	71.90	78.06	0.1471	8.3279
ReDet	72.69	80.03	**0.0882**	6.2136
RetinaNet	63.48	66.69	0.2050	11.1988
RetinaNet + GWD	65.11	66.55	0.1932	10.5551
RetinaNet + KLD	65.94	66.77	0.1941	11.2326
RetinaNet + ProbIoU	65.69	66.04	0.1979	10.4626
RoI Transformer	74.26	79.73	0.0942	5.6810
S2ANet	74.31	79.50	0.1090	6.8681

5.4. Biological Applications of OBB Cell Detection

Confluence estimation: The confluence MRE values considering all images in the test set for all evaluated methods are shown in the third column of Table 3. Two models (ReDet and RoI Transformer) presented an error below 10%, whereas FCOS produced the worst results with errors ∼30%. It is interesting to note that detectors with the highest AP do not necessarily produce the smallest confluence

[8] We used a Python package available at: https://pypi.org/project/kneed/

errors. In fact, Oriented RepPoints was the best detector (in both CAw and CAg strategies), but ranked 5th for confluence estimation. For the sake of comparison, the MRE considering all human annotators in the *Annotator's Comparison* split is 0.096, which is compatible with the results obtained by the best object detectors in the *Test* split.

Figure 8 shows the per-image confluence estimation produced by each method of the *Test* data split, along with the annotated GT value. We can observe that lower-density images (3–5, 14–24) tend to produce the smallest errors. The confluence errors tend to grow as the cell density increases, which might be expected: as noted in [46], high density and occlusions are still challenging scenarios for generic-purpose object detection. It is interesting to note that all detectors underestimated confluence in denser images, which might indicate the presence of false negatives. Lowering the detection score threshold could improve the results. Recall that Table 2 showed that using HBBs (in the CTMC dataset [2]) for the confluence measure would result in significant errors.

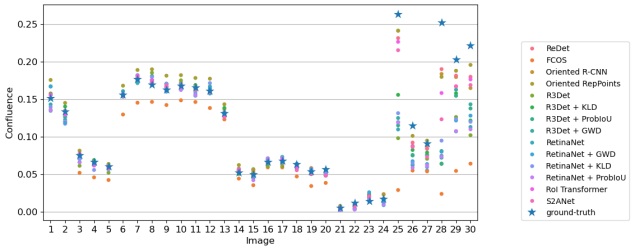


Figure 8. Cell confluence comparison for each image in the *Test* split.

Cell polarity: The polarity histogram χ^2 distance considering all images in the test set for all evaluated methods is shown in the fourth column of Table 3. According to this metric, Oriented RepPoints achieved the best result, closely followed by RoI Transformer and ReDet. For the sake of comparison, the mean χ^2 distance considering all human annotators in the *Annotator's Comparison* split is 7.34, which is compatible with the results obtained by several object detectors in the *Test* split.

Figure 9 shows a violin plot with the GT polarity histograms (interpolated with kernel density estimation) for each image in the test set, along with the corresponding results produced by the best method (Oriented RepPoints). As we can observe, the GT and estimated distributions are visually similar for most images. We can also note that images 21–30 present particularly long-tailed polarity distributions, indicating cells with large aspect ratios. Finally, we highlight that the polarity measures can only be performed in datasets containing angular information, so benchmarks such as [15] and [2] could not be used for these applications.

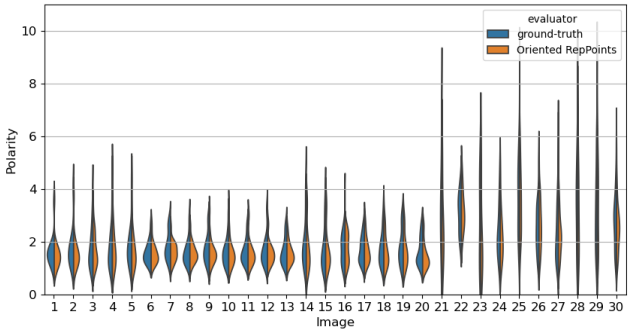


Figure 9. Cell polarity histogram (GT vs. Oriented RepPoints) for all images in the *Test* split.

6. Conclusions

This work introduced an Oriented Cell Dataset (OCD) containing bright-field microscopy images with OBB and category (elongated or round) annotations for cancerous human cells. A subset of the data contains labels from five different experts, allowing an Inter Annotator Analysis (IAA) suggesting $T = 0.5$ as a suitable IoU acceptance threshold for evaluating OBB object detectors in this dataset.

To corroborate our claim that OBBs and OEs are adequate representations for cell shapes, we presented a comparison of different annotation formats with the segmentation masks in the popular CTC dataset [23]. Our analysis showed that OBBs and OEs show considerable improvement over HBBs and AAEs, while being *much cheaper to annotate than full segmentation masks*. Moreover, the use of OBBs allows the automation of applications that require cell polarity measure [30], which is not possible in cases where there is no angular information.

Finally, we trained 14 SOTA OBB detectors in the OCD dataset and evaluated the results in terms of generic object detection (class-aware and class-agnostic AP values) and two important problems in cancer biology: cell confluence and polarity estimation. Our results indicated that the best model in terms of AP and polarity estimation was Oriented RepPoints, whereas the best one for confluence estimation was ReDet. However, we emphasize that such findings were based on the *Test* split, which contains annotations from a single annotator, and it could differ if compared to multiple human annotations. In fact, the metrics (MRE and χ^2) achieved by the best object detectors in the *Test split* are compatible with human variability.

Acknowledgements: This study was partially funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and FAPERGS (project RITE CIARS, 22/2551-0000390-7).

References

[1] Deep learning for computational cytology: A survey. *Medical Image Analysis*, 84:102691, 2023. 1, 2

[2] Samreen Anjum and Danna Gurari. CTMC: Cell tracking with mitosis detection dataset challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 982–983, 2020. 2, 3, 8

[3] Steven Busschots, Sharon O'Toole, John J O'Leary, and Britta Stordal. Non-invasive and non-destructive measurements of confluence in cultured adherent cell lines. *MethodsX*, 2:8–13, 2015. 1

[4] 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

[5] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu. Learning roi transformer for oriented object detection in aerial images. In *CVPR*, pages 2844–2853, 2019. 2

[6] 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. 7

[7] Eduardo C Filippi-Chiela, Manuel M Oliveira, Bruno Jurkovski, Sidia Maria Callegari-Jacques, Vinicius Duval da Silva, and Guido Lenz. Nuclear morphometric analysis (nma): screening of senescence, apoptosis and nuclear irregularities. 2012. 1

[8] Juliano T Freitas, Ivan Jozic, and Barbara Bedogni. Wound healing assay for melanoma cell migration. *Melanoma: Methods and Protocols*, pages 65–71, 2021. 1, 3, 6

[9] Zonghao Guo, Chang Liu, Xiaosong Zhang, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In *CVPR*, pages 8792–8801, 2021. 3

[10] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 2, 5

[11] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. 2, 5

[12] Junya Hayashida, Kazuya Nishimura, and Ryoma Bise. Consistent cell tracking in multi-frames with spatio-temporal context by object-level warping loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1727–1736, 2022. 1

[13] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007. 6

[14] Yves Jossin. Molecular mechanisms of cell polarity in a range of model systems and in migrating neurons. *Molecular and Cellular Neuroscience*, 106:103503, 2020. 6

[15] Dai Fei Elmer Ker, Sungeun Eom, Sho Sanami, Ryoma Bise, Corinne Pascale, Zhaozheng Yin, Seung-il Huh, Elvira Osuna-Highley, Silvina N Junkers, Casey J Helfrich, et al. Phase contrast time-lapse microscopy datasets with automated and manual cell tracking annotations. *Scientific data*, 5(1):1–12, 2018. 2, 3, 8

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. pages 4015–4026, 2023. 1

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

[18] 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. 7

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

[20] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5, 6

[21] Simona Martinotti and Elia Ranzato. Scratch wound healing assay. *Epidermal cells: methods and protocols*, pages 225–229, 2020. 1, 3, 6

[22] Martin Maška, Vladimír Ulman, Pablo Delgado-Rodriguez, Estibaliz Gómez-de Mariscal, Tereza Nečasová, Fidel A Guerrero Peña, Tsang Ing Ren, Elliot M Meyerowitz, Tim Scherr, Katharina Löffler, et al. The cell tracking challenge: 10 years of objective benchmarking. *Nature Methods*, 20(7):1010–1020, 2023. 2, 3

[23] Martin Maška, Vladimír Ulman, David Svoboda, Pavel Matula, Petr Matula, Cristina Ederra, Ainhoa Urbiola, Tomás España, Subramanian Venkatesan, Deepak MW Balak, et al. A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 30(11):1609–1617, 2014. 1, 2, 3, 4, 6, 8

[24] Jeffri Murrugarra-Llerena, Lucas N Kirsten, and Claudio R Jung. Can we trust bounding box annotations for object detection? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4813–4822, 2022. 4

[25] 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. 3, 5

[26] Joseph Nassar, Viveca Pavon-Harr, Marc Bosch, and Ian McCulloh. Assessing data quality of annotations with krippendorff alpha for applications in computer vision. *arXiv preprint arXiv:1912.10107*, 2019. 4

[27] Tran Thien Dat Nguyen, Hamid Rezatofighi, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustwor-

thy are performance evaluations for basic vision tasks? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4, 5

[28] 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. 7

[29] R. Paveley, N. Mansour, I. Hallyburton, A. Guidi, I. Gilbert, A. Hopkins, and Q. Bickle. Whole organism high-content screening by label-free, image-based bayesian classification for parasitic diseases. *Plos Neglected Tropical Diseases*, 6:e1762, 2012. 1

[30] Myriam Reffay, Laurence Petitjean, Sylvie Coscoy, Erwan Grasland-Mongrain, Francois Amblard, Axel Buguin, and Pascal Silberzan. Orientation and polarity in collectively migrating cell structures: statics and dynamics. *Biophysical journal*, 100(11):2566–2575, 2011. 1, 6, 8

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2

[32] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *International conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011. 7

[33] Andrew Oliveira Silva, Eloisa Dalsin, Giovana Ravizzoni Onzi, Eduardo Cremonese Filippi-Chiela, and Guido Lenz. The regrowth kinetic of the surviving population is independent of acute and chronic responses to temozolomide in glioblastoma cell lines. *Experimental Cell Research*, 348(2):177–183, 2016. 3

[34] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019. 2, 5

[35] Kaixuan Hu Jianke Zhu Wentong Li, Yijie Chen. Oriented reppoints for aerial object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5

[36] 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. 3

[37] 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. 5

[38] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3520–3529, October 2021. 2, 5

[39] 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*, pages 15819–15829, 2021. 2

[40] Xue Yang, Qingqing Liu, Junchi Yan, Ang Li, Zhiqiang Zhang, and Gang Yu. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612*, 2(4), 2019. 2, 5

[41] Xue Yang, Junchi Yan, Ming Qi, Wentao Wang, Zhang Xiaopeng, and Tian Qi. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International Conference on Machine Learning (ICML)*, 2021. 3, 5

[42] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *arXiv preprint arXiv:2106.01883*, 2021. 3, 5

[43] Xue Yang, Gefan Zhang, Xiaojiang Yang, Yue Zhou, Wentao Wang, Jin Tang, Tao He, and Junchi Yan. Detecting rotated objects as gaussian distributions and its 3-D generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[44] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. IoU loss for 2D/3D object detection. In *International Conference on 3D Vision*, pages 85–94, 2019. 3

[45] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, Wenwei Zhang, and Kai Chen. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 5

[46] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 1, 8