TU Delft

Delft University of Technology

Document Version
Final published version

Licence
Dutch Copyright Act (Article 25fa)

Citation (APA)
Strafforello, O., Kayhan, O. S., Inel, O., Schutte, K., & van Gemert, J. (2025). Aligning Object Detector Bounding Boxes with Human Preference. In A. Del Bue, C. Canton, J. Pont-Tuset, & T. Tommasi (Eds.), *Computer Vision – ECCV 2024 Workshops, Proceedings* (pp. 151-168). (Lecture Notes in Computer Science; Vol. 15634 LNCS). Springer. https://doi.org/10.1007/978-3-031-92591-7_10

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Aligning Object Detector Bounding Boxes with Human Preference

Ombretta Strafforello[1,2(✉)], Osman Semih Kayhan[1,3], Oana Inel[4], Klamer Schutte[2], and Jan van Gemert[1]

[1] Delft University of Technology, Delft, The Netherlands
{o.strafforello,o.s.Kayhan,j.c.vangemert}@tudelft.nl
[2] TNO, The Hague, The Netherlands
klamer.schutte@tno.nl
[3] Haga Teaching Hospital, The Hague, The Netherlands
[4] University of Zurich, Zurich, Switzerland
inel@ifi.uzh.ch

Abstract. Previous work shows that humans tend to prefer large bounding boxes over small bounding boxes with the same IoU. However, we show here that commonly used object detectors predict large and small boxes equally often. In this work, we investigate how to align automatically detected object boxes with human preference and study whether this improves human quality perception. We evaluate the performance of three commonly used object detectors through a user study (N = 123). We find that humans prefer object detections that are upscaled with factors of 1.5 or 2, even if the corresponding AP is close to 0. Motivated by this result, we propose an asymmetric bounding box regression loss that encourages large over small predicted bounding boxes. Our evaluation study shows that object detectors fine-tuned with the asymmetric loss are better aligned with human preference and are preferred over fixed scaling factors. A qualitative evaluation shows that human preference might be influenced by some object characteristics, like object shape.

Keywords: object detectors · human preference · box regression loss

1 Introduction

Object detectors identify and localize objects in an image. We focus on the common setting where detections are presented to a human by drawing a bounding box around the objects. In this paper, we evaluate how to best present object detections to humans, which is paramount for all applications that rely on showing detection to humans, such as visual inspection [32,37,46], anomaly detection [5,18,36], or medical imaging [38,43]. Previous work showed that humans prefer larger over smaller boxes with the same localization error [62]. This was concluded in an online study with a fully controlled setup, where ground truth bounding boxes are precisely matched to the localization error. However, it is not directly clear if this controlled setting translates to the real world, where

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Del Bue et al. (Eds.): ECCV 2024 Workshops, LNCS 15634, pp. 151–168, 2025.
https://doi.org/10.1007/978-3-031-92591-7_10

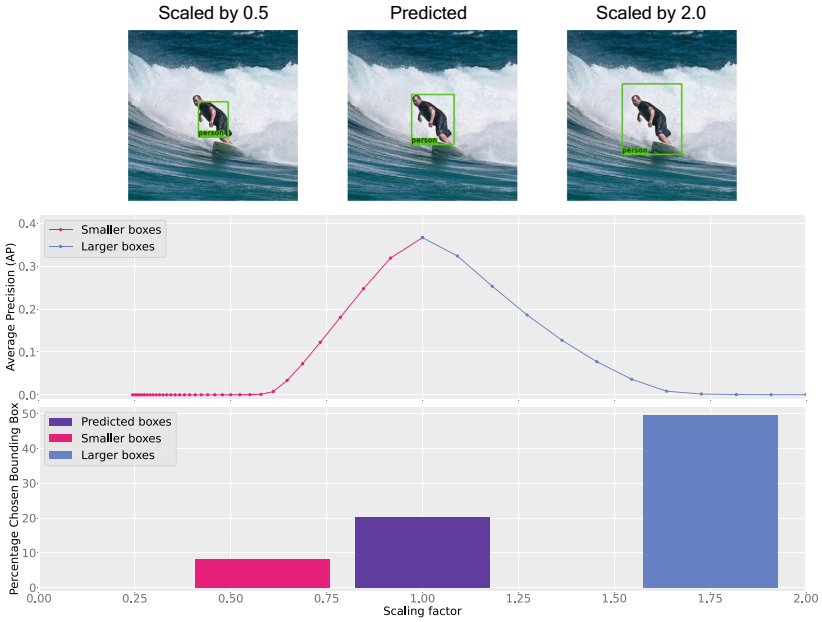


Fig. 1. Scaling the predicted bounding box of Faster R-CNN [53] on the COCO [41] validation set. Average Precision (AP) (*top*) versus human preference (*bottom*). A scaling factor of 1.0 corresponds to the original bounding box size. Upscaling and downscaling the size of the bounding boxes severely deteriorates AP. However, our study shows that humans prefer larger bounding boxes, even if they give nearly 0 AP.

object detector outputs are imperfect. In this work, we extend [62] to real-world settings and real object detectors, which is important for reproducibility, and realistic, practical applications of scientific results.

Object detectors, such as two-stage [10,53], single stage [40,44,50], anchor-less [19,35,73], and transformers-based detectors [7,12,16,75] minimize a classification loss and a localization loss for bounding box fitting. The localization loss is symmetric for errors in bounding box size: a predicted box that is 10% too large will give the same loss as a box that is 10% too small. Here, we investigate how this symmetry affects human perception of object detections.

Object detectors are typically evaluated using average precision (AP) [20, 29,34], which depends on the accuracy of the object classification and of the bounding box localization, as measured by the Intersection over Union (IoU) with the ground truth box. We are not the first to reconsider object detection evaluation [13,22,30], yet, those works all assume that a perfect-fitting bounding box is best. In contrast, here we investigate if *a perfect-fitting bounding box may not be the best box* for presenting detections to humans.

We make the following contributions: *(1)* We analyze three popular object detectors and find that they predict small and large bounding boxes equally often. *(2)* We analyze how humans perceive the predictions of the object detec-

tors focusing on the bounding box size. As shown in Fig. 1, we find that humans prefer upscaled object detections, even with corresponding AP close to 0. *(3)* We propose an asymmetric loss function that favors the prediction of large over small boxes. Our evaluation shows that fine-tuning with the asymmetric loss better aligns object detections with human preference. All our collected data, analyses, and code are available on GitHub[1].

2 Related Work

2.1 Presenting Object Detections to Humans

We take a nuanced view on evaluating object detection by identifying two distinct use-cases. Case 1: A bounding box is used as pre-processing for a follow-up algorithm such as instance segmentation [8,27,59], video object detection [14, 25,31], human pose estimation [11,64,69], action recognition [4], etc. Case 2: A bounding box is drawn on the image, and the full image is presented directly to a human, with relevant use-cases such as visual inspection [32,37,46], anomaly detection [5,18,36], medical imaging [38,43], etc. We argue that these two use-cases deserve different treatment. For case 1, where the bounding box is a pre-processing step, it is difficult to consider all possible follow-up algorithms, and a tightly fitting box around the object, as evaluated using IoU, seems reasonable. For case 2, however, the bounding box is the final end result and is shown to a human being. Case 2 allows directly evaluating the end result in user studies, to understand what humans actually prefer in their object detection. This is the focus of our paper.

2.2 Evaluating Object Detectors

All object detectors such as two-stage models [10,23,24,53], single stage approaches [40,44,50,51], pointwise/anchorless methods [19,35,73], and transformers-based detectors [7,12,16,75] are commonly evaluated [21,29,34,58] with mean average precision: the mean of the per-class average precision scores. Average precision (AP) is the area under the precision-recall curve, created by ranking all detections by confidence, and then checking if a detection is correct according to the ground truth. The correctness of a detection depends on the classification: if the assigned class label is wrong, the detection is wrong. A second criterion for correctness is that the location and size of the detection have sufficient overlap with the ground truth box. For determining the overlap, the Intersection over Union (IoU) score $\frac{B_p \cap B_{gt}}{B_p \cup B_{gt}}$ is used, where B_p is the predicted bounding box, and B_{gt} is the ground truth bounding box. The location of a detection is correct if the IoU score is higher than a certain threshold, typically 0.5 or higher [21,41]. Usually, the reported AP corresponds to a specific IoU threshold, such as 0.50 (AP50), or the average across several IoU thresholds, such as AP@[0.5 : 0.95].

[1] https://github.com/ombretta/humans-vs-detectors.

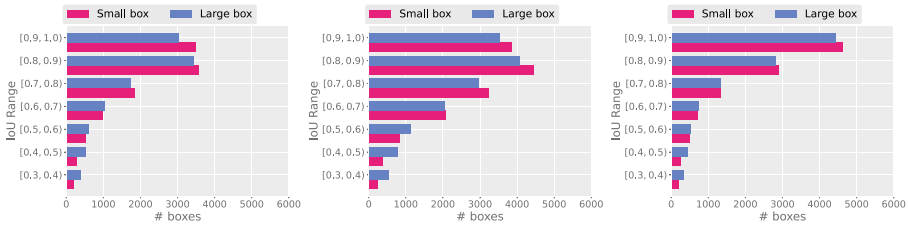
We are not the first to consider object detection evaluation [13,22,30,47, 49,60], yet, those works all assume that a predicted bounding box perfectly overlapping with the ground truth bounding box is best. In contrast, we here challenge the view that a best fitting bounding box is always best for presenting detections to humans. We base our challenge on the work of Strafforello *et al.* [62] who show in precisely controlled experiments on ground truth boxes that humans prefer larger boxes over smaller boxes. In this paper, we investigate the practical ramifications of Strafforello *et al.* [62] by aligning real-world object detectors with human preference.

2.3 Optimizing Object Detectors

Object detectors are typically optimized using an object classification loss and a bounding box regression loss for accurate localization, by aligning the IoU of the predicted box with the ground truth box. The regression loss, usually an L2 [52] or smoothed L1 [45,53] function, forces the box coordinates to be as close a possible to the ground truth, where the IoU is often optimized as an additional loss term [53]. Previous work proposed novel object detector losses to improve the accuracy, measured in AP. Examples include using the Absolute size IoU (AIoU) [65] and the SCALoss [72]. Other work designed a new loss term to achieve computational efficiency [3]. In our paper, we propose a simple asymmetric regression loss function that enhances the performance of object detectors with respect to human quality judgments. Previous work used asymmetric loss in Bayesian estimation [6] and for classification [54]. To the best of our knowledge, we are the first to use an asymmetric loss for bounding box regression.

2.4 Human Annotations for Object Detection

The adoption of crowdsourcing platforms such as Amazon Mechanical Turk [1] or Prolific [2] facilitated the collection of large training and testing datasets for computer vision tasks [9,17,33,42,57,71], in contrast to using in-house annotators [20,70]. For object detection, crowdsourcing studies are extensively used to draw bounding boxes around objects that appear in images [61,74] and videos [66] and to draw the precise shape of the object [57,71]. To eliminate the need for clustering or averaging several bounding boxes for the same object, in [55,63], the authors proposed a three-step workflow, where one annotator performs one step: *(1)* draws a bounding box around an object; *(2)* validates the drawn bounding box and *(3)* decides whether there are still objects that need to be annotated in the image. These steps are repeated until all objects in an image are annotated with bounding boxes. Experiments in which the crowd validates object detections showed that annotators tend to be lenient when validating bounding boxes, *i.e.*, bounding boxes with IoU < 0.5 are still accepted [48]. Furthermore, analyses performed in [56] suggest that to efficiently and accurately localize all objects in an image, several crowdsourcing tasks are needed, such as verifying box correctness, verifying object presence, or naming the object.



(a) Faster R-CNN (b) RetinaNet (c) Cascade Mask R-CNN

Fig. 2. Amount of large and small bounding boxes are predicted by three object detectors on the MS COCO dataset, for seven IoU intervals, ranging from 0.3 to 1.0. For all three detectors, with higher IoU thresholds more small than large boxes are detected.

3 Do Humans Prefer Larger Detections?

Previous work shows that, for equal IoU, humans prefer too large boxes over too small boxes [62]. Here, we evaluate if this has practical consequences for real object detectors. We use three popular object detectors pretrained on MS COCO: Faster R-CNN [53], RetinaNet [40], and Cascade Mask R-CNN with ResNet-50 [26] + Feature Pyramid Network [39] backbone [10,28] all implemented in the Detectron2 library [68].

3.1 Do Real Detectors Predict Too Large or Too Small Boxes?

If real object detectors tend to predict too large bounding boxes, then they are already well aligned with human preference. Thus, we investigate the relative size of the predicted bounding boxes with respect to the ground truth bounding box: A *small box* has a smaller predicted area, and a *large box* has a larger predicted area. We analyze predictions on the MS COCO validation set and count the occurrences of small and large boxes.

An overview of the distribution of the predicted bounding boxes over various IoU intervals is shown in Fig. 2. For all object detectors that we examined, there is no statistically significant difference in the number of occurrences of large and small bounding boxes. This holds for small, medium, and large objects. However, for low IoU ranges, i.e., IoU $\in [0.3, 0.6)$ for Faster R-CNN and RetinaNet and IoU $\in [0.3, 0.5)$ for Cascade Mask R-CNN, large bounding boxes are more frequent than small ones. This is due to random large bounding boxes being more likely to partially overlap with the ground truth, compared to random small boxes. Considering intermediate IoU ranges, like IoU $\in [0.6, 0.7]$, the number of occurrences of small and large boxes is not in line with the human preference found in Strafforello *et al.* [62]. That is, where humans would choose a large box over a small box with approximately 70% chance, an object detector would predict small or large with nearby equal probability.

We conclude that real object detectors generally do not predict too large boxes more often than too small boxes, and thus seem not well-aligned with

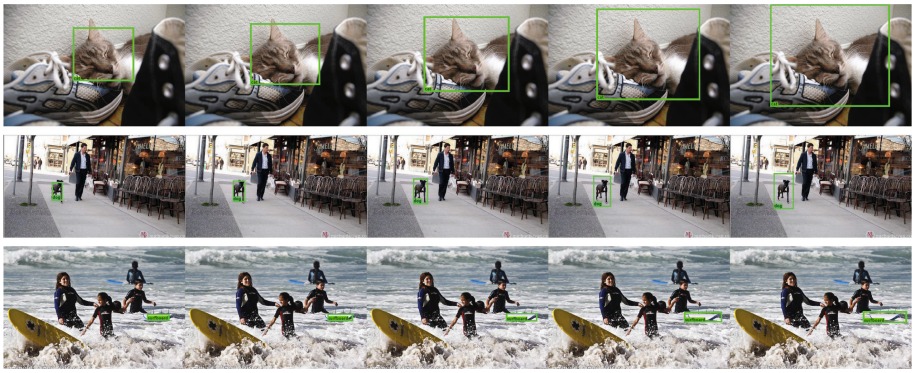


Fig. 3. Scaling the model detections. Example of a bounding box predicted for a large object (first row), a medium object (second row) and a small object (third row) with Faster R-CNN (3rd column) and its scaled versions. In the left two images, the area of the bounding box is reduced by a scaling factor of, respectively, 0.5 and 0.67, whilst in the right two images the box area is increased by a factor of 1.5 and 2.

Table 1. AP (*i.e.*, AP@[0.5 : 0.95]) and AP50 (%) calculated for the predictions of three detectors on the MS COCO validation set and for the predicted boxes scaled with different scaling factors. Scaling the predicted boxes reduces the AP scores drastically.

Scaling factor	Faster R-CNN		RetinaNet		Cascade R-CNN	
	AP	AP50	AP	AP50	AP	AP50
0.50	0.0	0.1	0.1	0.4	0.0	0.1
0.67	5.1	37.1	5.4	38.2	5.6	40.7
1.00	36.7	54.1	37.4	56.7	39.6	53.7
1.50	5.5	38.4	6.2	41.3	5.7	41.3
2.00	0.0	0.2	0.3	1.3	0.0	0.2

human preference. In the following, we will investigate what this means for human quality judgments of real object detectors.

3.2 For Real Object Detectors, Do Humans Prefer Too Large Boxes or Too Small Boxes?

Given that, for the same IoU, humans prefer larger boxes and real object detectors do not tend to predict too large boxes, here we evaluate how humans judge re-scaled boxes. We do this through a user study, where we ask participants to evaluate five scaling factors, determined by scaling up or down the area of the predicted boxes with a factor of 1.5 and 2.0: $\{0.5, 0.67, 1.0, 1.5, 2.0\}$. Large bounding boxes are cropped to not exceed the image boundaries. Examples of bounding box scaling for a large and a small object are shown in Fig. 3. We

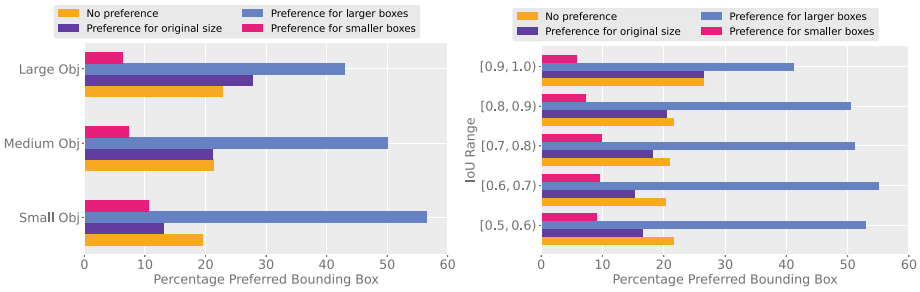
refer to this study as *Scaling Preference*. We ask the participants to choose the boxes they believe best identify a specific object in an image. The interface used in the user study allows the participants to select multiple options if they cannot determine a single best one. We use six random images selected from the MS COCO validation set per each combination between object size (*small*, *medium*, *large*) and IoU range. We select five IoU ranges from $0.5 \leq \text{IoU} < 0.6$ to $0.9 \leq \text{IoU} < 1.0$ that correspond to true positive predictions, for a total of 90 images. We conduct this *Scaling Preference* study on Faster R-CNN, RetinaNet, and Cascade Mask R-CNN.

Scaling the detections of a well-performing object detector results in a slight change in appearance but a significant drop in AP. For a scaling of 1, the baseline AP is 36.7%, yet a scaling of 1.5 corresponds to a $\approx 86\%$ decrease in AP. For a scaling factor of 2.0, the AP is $\approx 0\%$. Even with a more lenient IoU threshold, the AP50 decreases rapidly with both upscaling and downscaling. As shown in Table 1, this behavior is consistent across the three object detectors.

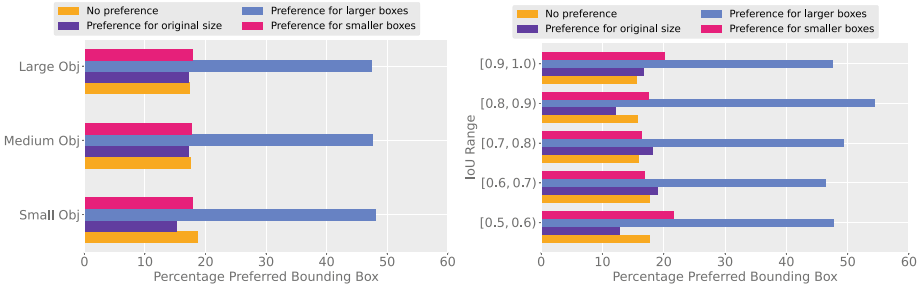
3.3 Results for the *Scaling Preference* study

Table 2 shows the number of participants and total number of judgments for the *Scaling Preference* study. We use the Cochran's Q test [15] to determine whether there are statistically significant differences in participants' preferences regarding box sizes. In addition, we apply the posthoc Dunn tests with Bonferroni correction [67] to find what are the scaling factors that result in significant differences in users' preferences. An overview of the results is provided in Fig. 4. We group the scaling choices into (i) "Preference for smaller boxes", if a user selected the box scaled with factor 0.67, the box scaled with factor 0.5 or both; (ii) "Preference for larger boxes", if a user selected the box scaled with factor 1.5, the box scaled with factor 2.0 or both; (iii) "Preference for original size" if a user selected only the bounding box predicted by the model and (iv) "No preference" for all the remaining combinations of selections. Larger bounding boxes are consistently selected more often than small bounding boxes and than the original bounding box size for all three object detectors. This holds for different object sizes (Fig. 4, left column), and IoU ranges (Fig. 4, right column).

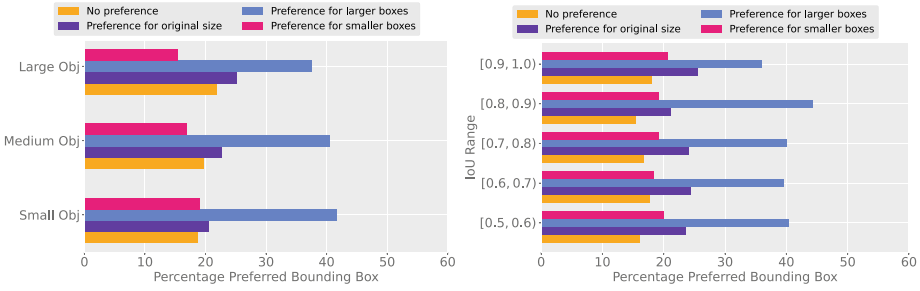
Despite the preference for larger boxes, we cannot find a statistically significant difference between the preference for upscaling factor 1.5 and upscaling factor 2.0. For Faster R-CNN, the preference for larger boxes is composed of 56.77% of selections of both boxes scaled with factor 1.5 and 2.0; of 19.30% of selections for scaling factor 1.5 and of 23.93% of selections for scaling factor 2.0. Here, Dunn's test shows no statistically significant difference between the preference for scaling factor 1.5 and scaling factor 2.0. This means that larger boxes are preferred, but there is no single best upscaling factor. One exception holds for the bounding boxes for small objects predicted with Faster R-CNN: in this case, scaling factor 2.0 is preferred over scaling factor 1.5 (Dunn's $\alpha \approx 0$). This preference is an indicator that, for small objects, scaling the bounding box with a large scaling factor, like 2.0, results in more satisfactory detections. A majority of votes for the largest box for small objects, albeit not statistically significant,



(a) Faster R-CNN



(b) RetinaNet



(c) Cascade Mask R-CNN

Fig. 4. Results from the *Scaling Preference* user study. The histograms show the percentage of preferred bounding box size per object category (S, M, L) and IoU range, from $0.5 \leq \text{IoU} < 0.6$ to $0.9 \leq \text{IoU} < 1.0$, for three object detectors. The plots indicate that humans significantly prefer larger boxes.

is observed for the other object detectors. It is noticeable how larger bounding boxes are preferred to bounding boxes predicted with high IoU. This indicates that, for representative images of the diverse MS COCO dataset, humans are likely to prefer bounding boxes larger than the ground truth bounding boxes.

Table 2. Overview of participants and their judgments in the scaling preference study.

	Faster R-CNN	RetinaNet	Cascade Mask R-CNN
Participants	39	36	48
Judgments	5400	5220	5632

4 Asymmetric Regression Loss to Encourage Larger Detections

We find that humans consistently prefer larger object detections, while object detectors predict large and small boxes equally often. We propose an asymmetric bounding box regression loss that encourages larger detections. Our asymmetric loss is obtained by a simple modification of the smooth L_1 localization loss function used in standard object detectors. We use the asymmetry term α to increase the loss value when the predicted area is smaller than the ground truth area and decrease the loss value when is larger. The asymmetric loss is given by

$$\text{Asymmetric } L_{1,\text{smooth}} = \begin{cases} \frac{1}{2\sqrt{\alpha}\beta}x^2, & \text{if } 0 \leq x < \beta \\ \frac{\sqrt{\alpha}}{2\beta}x^2, & \text{if } -\beta < x < 0 \\ \frac{1}{\sqrt{\alpha}}x - \frac{\beta}{2\sqrt{\alpha}}, & \text{if } x \geq \beta \\ -\sqrt{\alpha}x - \frac{\sqrt{\alpha}\beta}{2}, & \text{if } x \leq -\beta \end{cases} \tag{1}$$

The α represents the asymmetry term, β determines the standard smoothing interval in which the L_1 loss becomes quadratic, and x is the input to the loss function, which is the difference between the predicted height/width and the ground truth values, $x = x_{\text{pred}} - x_{\text{GT}}$. As shown in Fig. 5, the asymmetric loss is identical to the smooth L_1 loss when $\alpha = 1$. We use the asymmetric loss function for the regression of the boxes' height and width.

We fine-tune Faster R-CNN, RetinaNet, and Cascade R-CNN on MS COCO for 100k iterations with the asymmetric loss. As a result, the fine-tuned models are more likely to predict larger boxes over smaller boxes. Figure 6 shows the percentage of large detections for different α values. Similarly to the fixed scaling factors in the *Scaling Preference* study in Sect. 3, we observe a decrease in the AP with the increase of large detections. With $\alpha = 10$, we obtain 80% to 90% large predictions without compromising AP too much.

We measure the average size increase of the predicted bounding boxes compared to the ground truth. As shown in Fig. 7, increasing the α coefficient results in an increase of the average box size, for all three models and object sizes. The models fine-tuned with $\alpha = 10$ return detections scaled compared to ground truth, on average by factors 1.21 ± 0.24 for Faster R-CNN, 1.21 ± 0.25 for RetinaNet, and 1.19 ± 0.22 for Cascade R-CNN, while fine-tuning with $\alpha = 100$ results in average scaling of 1.41 ± 0.26 for Faster R-CNN, 1.34 ± 0.28 for RetinaNet, and 1.39 ± 0.24 for Cascade R-CNN. It is noticeable that the size of small objects' detections increases more than for medium and large objects. This is

mostly due to small bounding boxes having more opportunity for expansion in the image, while large objects' boxes are already close to the image boundaries.

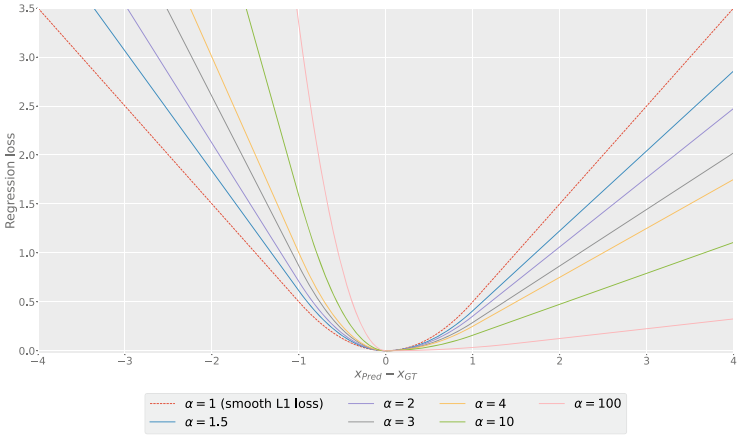


Fig. 5. Asymmetric smooth L_1 loss with different α. Larger bounding boxes are penalized less than smaller predicted boxes.

4.1 Does the Asymmetric L_1 Loss Lead to Detections Closer to the Human Preference?

We conduct a final user study to investigate whether adopting the asymmetric L_1 loss results in detections closer to human preference. In the evaluation study, we include the detections from the original pretrained Faster R-CNN ($\alpha = 1$), the detections from Faster R-CNN fine-tuned with α 10 and 100, and the detections scaled by a fixed factor 1.5, which was one of the preferred options in the *Scaling Preference* study Sect. 3. These values for α are chosen to have detection sizes that notably differ from the Faster R-CNN baseline (Fig. 7). We ask users to compare the four different detections for the same object and choose the one that, in their opinion, best identifies the object. We include 45 detections, equally sampled from the three object categories (*small*, *medium*, *large*). We conduct the study on Amazon Mechanical Turk [1] and collect 660 judgments.

The results are summarized in Table 3. The Cochran's Q test reveals statistically significant differences between the proportions of preferred object detections. The detections obtained by fine-tuning with asymmetric loss, $\alpha = 10$, are always the most preferred. This preference is statistically significant when con-

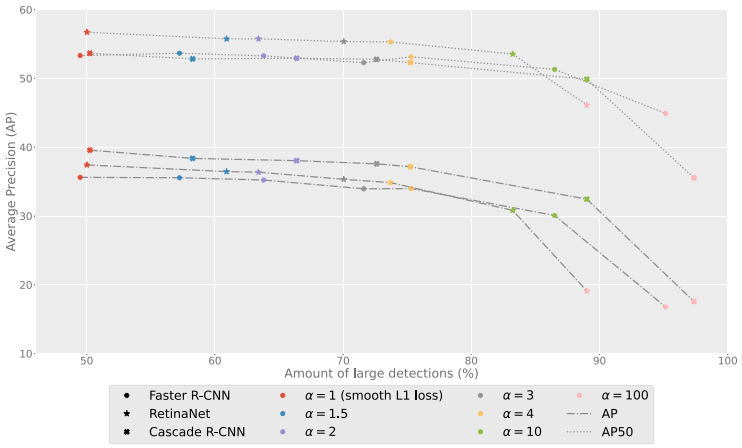


Fig. 6. Average Precision (AP) as a function of the amount of predicted boxes that are larger than the ground truth boxes. The percentage of large detections increases with the α parameter, while the AP decreases.

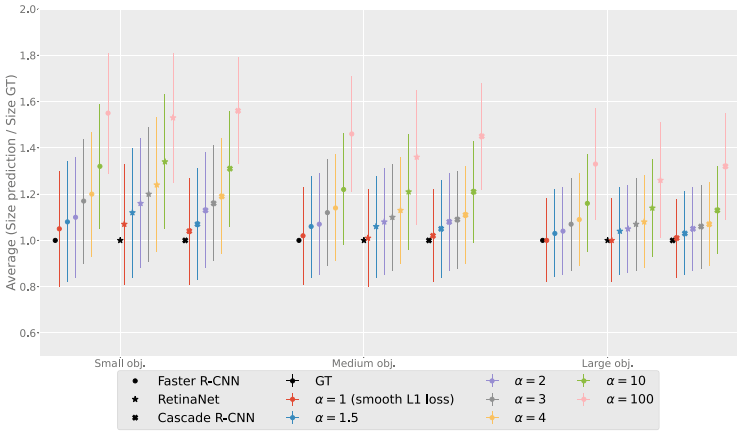


Fig. 7. Bounding box size increases after fine-tuning object detectors with the asymmetric smooth L_1 loss with parameter α.

sidering all object categories and small objects (Dunn's $\alpha \leq 0.001$). In the other cases, fine-tuning with $\alpha = 10$ is significantly more preferred than scaling with a fixed factor (Dunn's $\alpha \approx 0$), thus confirming the advantage of the asymmetric loss over fixed scaling.

The preference for the asymmetric loss over fixed scaling is likely due to the fixed scaling factor upscaling all boxes equally, irrespective of the object size. Conversely, using the asymmetric loss results in boxes upscaled more for small objects than for medium and large objects, as illustrated in Fig. 7. This

might lead to higher human preference, since large objects are already easily identifiable with a tighter box. In fact, as shown in Table 3, the fixed scaling 1.5 is almost never chosen for large objects. In addition, we observe that the most preferred option—asymmetric loss with $\alpha = 10$—leads to detections that are, on average, larger than the ground truth by a factor between 1.1 and 1.5. We hypothesize that the optimal scaling factor might lie within this range. Another potential reason why scaling by 1.5 is less preferred is that the fixed scaling strategy retains the aspect ratio of the original predicted box. This aspect ratio may not be optimal when upscaling the boxes. In contrast, the asymmetric loss function imposes fewer constraints on the aspect ratio.

Overall, fine-tuning the models with our asymmetric L_1 loss results in detections closer to human preference. We suggest adopting this loss when object detections are meant to be presented to humans.

Table 3. Users' preferred object detections (%), computed with Faster R-CNN fine-tuned with the asymmetric loss function or up-scaled with factor 1.5. Fine-tuning with $\alpha = 10$ is always the most preferred option.

Object cat.	# judgments	Chosen object detection (%)			
		$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	Scal. fact. 1.5
All	660	27.4	37.1	21.2	14.2
Small	229	17.0	31.9	26.6	24.5
Medium	215	27.9	39.5	16.7	15.8
Large	216	38.0	40.3	19.9	1.9

4.2 Qualitative Analysis of the Preferred Boxes

We manually analyze the results obtained from the user evaluation of the asymmetric loss and illustrate some representative examples in Fig. 8. We notice that the tight bounding boxes predicted by the Faster R-CNN baseline, namely, trained with $\alpha = 1$, are generally preferred for large objects, e.g., the *cat* in the first row. Preference for $\alpha = 1$ also occurs when there are multiple objects behind or in the proximity of the object of interest. In the image on row 2 of Fig. 8, larger bounding boxes partly include the *chair* behind the one of interest. In this case, tight boxes delineate better the subject of focus.

Slightly larger boxes, obtained by fine-tuning with our asymmetric loss, $\alpha = 10$, are preferred when small parts of the object are not contained in the tight bounding box predicted by Faster R-CNN (e.g., the *candle* on the birthday cake in Fig. 8, row 3), or partly covered by the box contour itself, like the ears and tail of the *cat* in row 4. We hypothesize that predicted tight bounding boxes leave out possible object protrusions, despite resulting in high AP. Presumably, humans prefer large boxes because they can include the whole object. Additionally, in

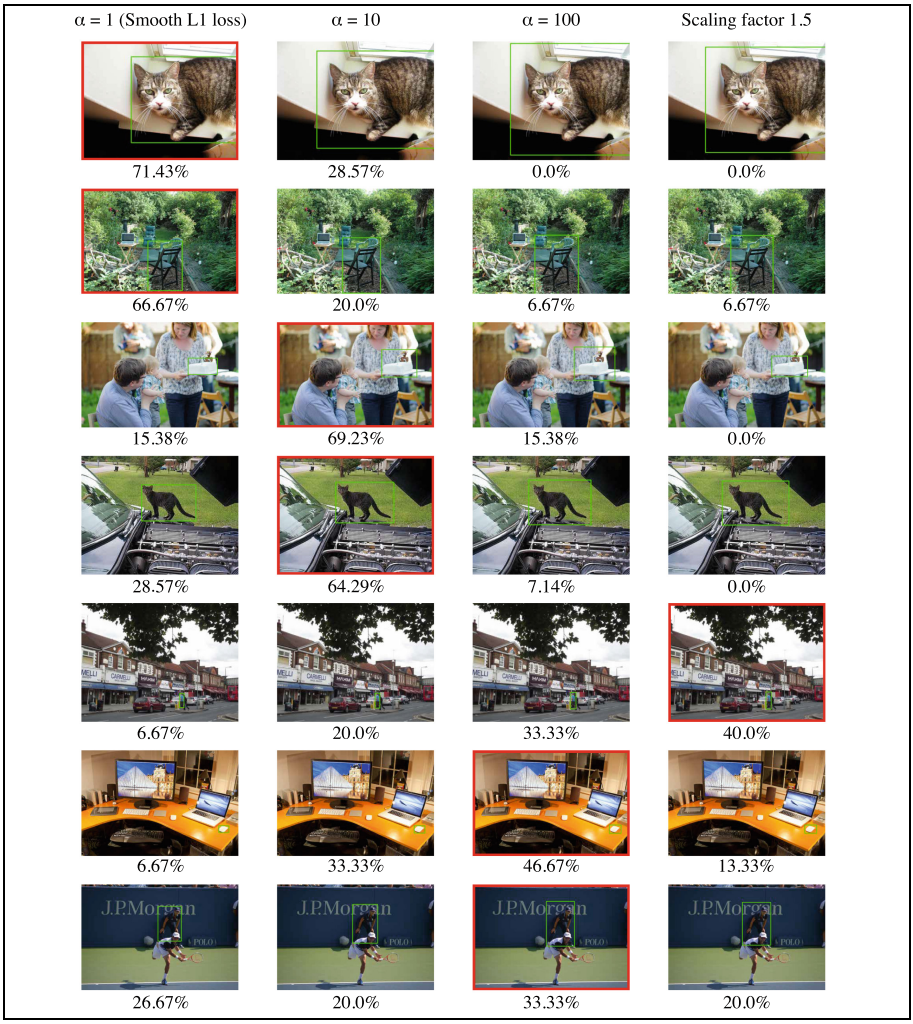


Fig. 8. Example of human preferences obtained from the user evaluation of the asymmetric loss. The columns show the percentage of users who prefer the bounding boxes obtained by the original pretrained Faster R-CNN ($\alpha = 1$), after fine-tuned with α 10 and 100, or scaling by a fixed factor 1.5. Generally, humans prefer tight boxes for large objects (first row) and when the object of interest overlaps with other objects (second row). Slightly larger boxes, obtained with asymmetric loss, $\alpha = 10$ or 100, are preferred when small parts of the object protrude outside too tight bounding boxes (e.g., the candle on the birthday cake, third row), or partly covered by the box line itself (fourth row). Large boxes ($\alpha = 1$ or scaling factor 1.5) are chosen for very small objects (fourth and fifth row). Finally, we found no preference when all bounding boxes are too visually similar (last row).

the presence of a uniform background (e.g., the green grass behind the cat in Fig. 8, row 4), humans are generally less concerned if the bounding box is slightly larger. Similarly, the asymmetric loss makes it more likely to include all the small protruding parts of the objects in the predicted boxes.

We observe that the preference for larger boxes, obtained by scaling with factor 1.5 or with the asymmetric loss $\alpha = 100$ occurs when the objects of interest are very small, e.g., the *mouse* and the *person* walking on the street (row 5 and 6, Fig. 8). Finally, in a few cases, the original Faster R-CNN detector, the detectors fine-tuned with the asymmetric loss or manually scaled result in very similar boxes, indistinguishable by a human eye. In this situation, we observe no clear human preference, as for the *person* in the last row in Fig. 8.

The qualitative analysis suggests that there exists a relationship between the object characteristics, especially size (already observed in [62]) and shape, and the preferred bounding box size. We leave the investigation of the factors that determine the user preference for future work.

5 Conclusion

Prior work [62] shows that humans prefer larger boxes in a fully controlled setup. In this paper, we confirm this result in practice, with real detectors. We evaluate the bounding boxes predicted by three popular object detectors. We find that the object detectors predict large and small bounding boxes equally often, therefore are not aligned with the human preference found in [62]. In addition, humans consistently prefer larger bounding boxes over the predicted boxes, even with AP approximately zero. Therefore, we recommend being careful with AP scores when object detectors are intended for human use: a high AP does not automatically correspond to high human preference.

It is noticeable how the preference occurs even for bounding boxes predicted with high IoUs: this suggests that humans are likely to prefer larger bounding boxes compared to tight ground truth bounding boxes.

We propose an asymmetric loss function that encourages detectors to predict large boxes more often than small boxes, without having to re-annotate the training images. Our user evaluation shows that fine-tuning with the asymmetric loss results in object detections more aligned with human preference. After qualitatively analyzing the results collected from our study, we hypothesize that the human preference is affected by the object characteristics, such as shape and size. For example, generally tight boxes are preferred for large objects, while larger boxes are preferred for small objects. Further investigation into these observations may be considered in the future.

Acknowledgements. This work is part of the research program Efficient Deep Learning (EDL), which is (partly) financed by the Dutch Research Council (NWO).

References

1. Amazon mechanical turk. https://www.mturk.com/. Accessed 20 Feb 2023
2. Prolific. https://www.prolific.co
3. Aswal, D., Shukla, P., Nandi, G.C.: Designing effective power law-based loss function for faster and better bounding box regression. Mach. Vis. Appl. **32**(4), 1–10 (2021). https://doi.org/10.1007/s00138-021-01206-5
4. Ballan, L., Strafforello, O., Schutte, K.: Long-term behaviour recognition in videos with actor-focused region attention. In: VISIGRAPP (5: VISAPP) (2021)
5. Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: CVPR. IEEE/CVF (2008)
6. Basu, A., Ebrahimi, N.: Bayesian approach to life testing and reliability estimation using asymmetric loss function. J. Stat. Plan. Inference **29**(1–2), 21–31 (1991)
7. Beal, J., Kim, E., Tzeng, E., Park, D.H., Zhai, A., Kislyuk, D.: Toward Transformer-Based Object Detection (2020)
8. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT: real-time instance segmentation. In: ICCV. IEEE/CVF (2019)
9. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: ActivityNet: a large-scale video benchmark for human activity understanding. In: CVPR (2015)
10. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: CVPR. IEEE/CVF (2018)
11. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1) (2019)
12. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
13. Chavali, N., Agrawal, H., Mahendru, A., Batra, D.: Object-proposal evaluation protocol is 'gameable'. In: CVPR. IEEE/CVF (2016)
14. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: CVPR. IEEE/CVF (2020)
15. Cochran, W.G.: The comparison of percentages in matched samples. Biometrika **37**(3/4) (1950)
16. Dai, Z., Cai, B., Lin, Y., Chen, J.: UP-DETR: unsupervised pre-training for object detection with transformers. In: CVPR, pp. 1601–1610. IEEE/CVF (2021)
17. Di Salvo, R., Giordano, D., Kavasidis, I.: A crowdsourcing approach to support video annotation. In: Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications. ACM (2013)
18. Doshi, K., Yilmaz, Y.: Fast unsupervised anomaly detection in traffic videos. In: CVPRW. IEEE/CVF (2020)
19. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: keypoint triplets for object detection. In: ICCV. IEEE/CVF (2019)
20. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV (2010)
21. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
22. Feng, D., et al.: Labels are not perfect: inferring spatial uncertainty in object detection. IEEE Trans. Intell. Transp. Syst. (2021)

23. Girshick, R.: Fast R-CNN. In: ICCV. IEEE (2015)
24. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. PAMI (2015)
25. Han, L., Wang, P., Yin, Z., Wang, F., Li, H.: Context and structure mining network for video object detection. IJCV (2021)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. IEEE/CVF (2016)
27. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV. IEEE/CVF (2017)
28. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR **abs/1703.06870** (2017). http://arxiv.org/abs/1703.06870
29. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 340–353. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_25
30. Hosang, J., Benenson, R., Dollár, P., Schiele, B.: What makes for effective detection proposals? IEEE Trans. Pattern Anal. Mach. Intell. **38**(4) (2015)
31. Jiao, L., et al.: New generation deep learning for video object detection: a survey. IEEE Trans. Neural Netw. Learn. Syst. (2021)
32. Kayhan, O.S., Vredebregt, B., van Gemert, J.C.: Hallucination in object detection–a study in visual part verification. In: ICIP. IEEE (2021)
33. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. IJCV (2017)
34. Kuznetsova, A., et al.: The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. IJCV (2020)
35. Law, H., Deng, J.: CornerNet: detecting objects as paired keypoints. In: ECCV. Springer (2018)
36. Li, X., Li, W., Liu, B., Liu, Q., Yu, N.: Object-oriented anomaly detection in surveillance videos. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018)
37. Li, Y., Trinh, H., Haas, N., Otto, C., Pankanti, S.: Rail component detection, optimization, and assessment for automatic rail track inspection. IEEE Trans. Intell. Transp. Syst. **15**(2) (2013)
38. Li, Z., Dong, M., Wen, S., Hu, X., Zhou, P., Zeng, Z.: CLU-CNNs: object detection for medical images. Neurocomputing **350**, 53–59 (2019)
39. Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. IEEE/CVF (2017)
40. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. IEEE/CVF (2017)
41. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
42. Lintott, C.J., et al.: Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. Monthly Notices of the Royal Astronomical Society (2008)
43. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42** (2017)
44. Liu, W., et al.: SSD: single shot multibox detector, vol. 9905, pp. 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2

45. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
46. Mery, D., Katsaggelos, A.K.: A logarithmic x-ray imaging model for baggage inspection: simulation and object detection. In: CVPRW. IEEE/CVF (2017)
47. Padilla, R., Netto, S.L., da Silva, E.A.: A survey on performance metrics for object-detection algorithms. In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE (2020)
48. Papadopoulos, D.P., Uijlings, J.R.R., Keller, F., Ferrari, V.: We don't need no bounding-boxes: training object class detectors using only human verification. In: CVPR. IEEE/CVF (2016)
49. Prasad, D.K., Dong, H., Rajan, D., Quek, C.: Are object detection assessment criteria ready for maritime computer vision? IEEE Trans. Intell. Transp. Syst. **21**(12) (2019)
50. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR. IEEE/CVF (2016)
51. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
52. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: CVPR. IEEE/CVF (2017)
53. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015)
54. Ridnik, T., et al.: Asymmetric loss for multi-label classification. In: CVPR. IEEE/CVF (2021)
55. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. IJCV (2015)
56. Russakovsky, O., Li, L.J., Fei-Fei, L.: Best of both worlds: human-machine collaboration for object annotation. In: CVPR. IEEE/CVF (2015)
57. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. IJCV (2008)
58. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill Inc, USA (1986)
59. Shen, Y., et al.: Parallel detection-and-segmentation learning for weakly supervised instance segmentation. In: ICCV. IEEE/CVF (2021)
60. Sobti, A., Mavi, V., Balakrishnan, M., Arora, C.: VmAP: a fair metric for video object detection. In: Proceedings of the 29th ACM International Conference on Multimedia (2021)
61. Song, S., Zhang, L., Xiao, J.: Robot in a room: toward perfect object recognition in closed environments. CoRR, abs/1507.02703 (2015)
62. Strafforello, O., Rajasekar, V., Kayhan, O.S., Inel, O., van Gemert, J.C.: Humans disagree with the IoU for measuring object detector localization error. In: ICIP. IEEE (2022)
63. Su, H., Deng, J., Fei-Fei, L.: Crowdsourcing annotations for visual object detection. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
64. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. IEEE/CVF (2019)
65. Tian, D., Han, Y., Wang, S., Chen, X., Guan, T.: Absolute size IoU loss for the bounding box regression of the object detection. Neurocomputing **500**, 1029–1040

(2022). https://doi.org/10.1016/j.neucom.2022.06.018. https://www.sciencedirect.com/science/article/pii/S0925231222007378
66. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. IJCV (2013)
67. Weisstein, E.W.: Bonferroni correction (2004). https://mathworld.wolfram.com/
68. Wu, Y., Kirillov, A., Massa, F., Lo, W., Girshick, R.: Detectron2 (2019). https://github.com/facebookresearch/detectron2
69. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 472–487. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_29
70. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on Computer Vision and Pattern Recognition (2010)
71. Yuen, J., Russell, B., Liu, C., Torralba, A.: LabelMe video: building a video database with human annotations. In: CVPR. IEEE/CVF (2009)
72. Zheng, T., Zhao, S., Liu, Y., Liu, Z., Cai, D.: SCALoss: side and corner aligned loss for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 3535–3543 (2022)
73. Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: CVPR. IEEE/CVF (2019)
74. Zhu, X., Vondrick, C., Ramanan, D., Fowlkes, C.C.: Do we need more training data or better models for object detection?. In: BMVC, vol. 3. Citeseer (2012)
75. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)