Machine Vision and Applications (2024) 35:49
https://doi.org/10.1007/s00138-024-01527-1

RESEARCH

# The improvement of ground truth annotation in public datasets for human detection

Sotheany Nou[1] · Joong-Sun Lee[2] · Nagaaki Ohyama[2] · Takashi Obi[1,2]

Received: 31 May 2023 / Revised: 16 February 2024 / Accepted: 4 March 2024 / Published online: 8 April 2024
© The Author(s) 2024

**Abstract**

The quality of annotations in the datasets is crucial for supervised machine learning as it significantly affects the performance of models. While many public datasets are widely used, they often suffer from annotations errors, including missing annotations, incorrect bounding box sizes, and positions. It results in low accuracy of machine learning models. However, most researchers have traditionally focused on improving model performance by enhancing algorithms, while overlooking concerns regarding data quality. This so-called model-centric AI approach has been predominant. In contrast, a data-centric AI approach, advocated by Andrew Ng at the DATA and AI Summit 2022, emphasizes enhancing data quality while keeping the model fixed, which proves to be more efficient in improving performance. Building upon this data-centric approach, we propose a method to enhance the quality of public datasets such as MS-COCO and Open Image Dataset. Our approach involves automatically retrieving missing annotations and correcting the size and position of existing bounding boxes in these datasets. Specifically, our study deals with human object detection, which is one of the prominent applications of artificial intelligence. Experimental results demonstrate improved performance with models such as Faster-RCNN, EfficientDet, and RetinaNet. We can achieve up to 32% compared to original datasets in the term of mAP after applying both proposed methods to dataset which is transformed the grouped of instances to individual instance. In summary, our methods significantly enhance the model's performance by improving the quality of annotations at a lower cost with less time than manual improvement employed in other studies.

**Keywords** Ground truth annotations · Annotations errors · Public datasets · Human detection · Data-centric AI · Model-centric AI

## 1 Introduction

The human detection model is a prominent application of artificial intelligence for identifying the location of human objects in images or videos. This model can be developed using machine learning techniques that focus on learning algorithms from a given datasets. It is important to note that the performance of a model depends on both data and algorithms employed. Traditionally, researchers have been primarily improving algorithms or models to enhance performance while keeping the data fixed from the original sources. This approach is known as the "Conventional Model-Centric" approach, commonly used in the development of modern neural networks [1]. This method is commonly used to produce or invent new network models. For instance, a convolutional neural network with 50 layers deep (ResNet-50) is improved by applying Recursive Feature Pyramid (RFP) and Switchable Atrous Convolution (SAC). This combination of RFP and SAC on ResNet-50 improved 7.7% compared to Hybrid Task Cascade [2]. PP-YOLO used PaddlePaddle to improve YOLOv3 model performance on raw datasets [3]. As well as a Google research team proposed Bi-directional Feature Pyramid Network (BiFPN)

✉ Sotheany Nou
nou.s.aa@m.titech.ac.jp

Joong-Sun Lee
j-lee@isl.titech.ac.jp

Nagaaki Ohyama
yama@isl.titech.ac.jp

Takashi Obi
obi.t.aa@m.titech.ac.jp

[1] Department of Information and Communication Engineering, Tokyo Institute of Technology, Tokyo, Japan

[2] Institute of Innovative Research, Tokyo Institute of Technology, Tokyo, Japan

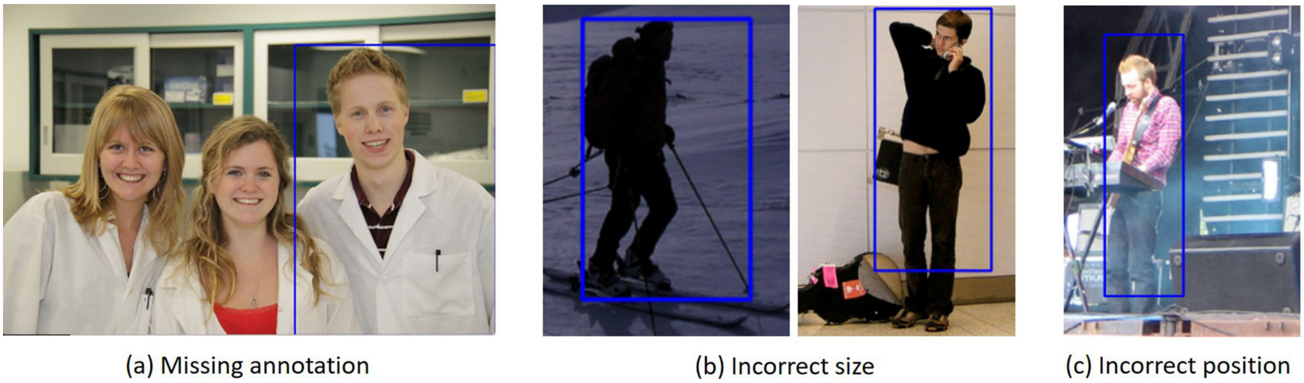


**Fig. 1** Inappropriate annotations in MS-COCO and Open Image Dataset. **a** Shows missing annotation in the image where only one person is annotated while the others are missed. **b** Consists of incorrect size of the bounding box in person class. The left bounding box includes the ski pole making the size unsuitably bigger. The right bounding box does not enclose the full body. **c** Displays the case that bounding box does not draw in the correct location which is shifted up

and compound scaling on the EfficientNet backbones to improve the object detector accuracy, named EfficientDet [4]. These researchers tended to modify the model for improving the performance while they kept MS-COCO 2017 dataset unchanged. However, it becomes more efficient even though algorithm is held fixed and focus on the data quality improvement. This approach is known as “Data-Centric AI" which is the discipline of systematically engineering used to build an AI system [1]. The concept of data-centric AI, the datasets can be improved by applying data augmentation, generation, collection, and annotation correction [5]. Even so, data collection will be expensive and time-consuming. There are public datasets MS-COCO [6], Pascal VOC [7], ImageNet [8] and Open Images Dataset [9] which are used by many studies to solve numerous encounter problems. However, some studies [5, 10, 11] mentioned that the data annotation process is still partially a manual task. It was performed by humans with various experiences which resulted in noisy annotations such as inaccurate bounding boxes and incorrect class labels. Based on the study [10], the most common object annotation errors are the missing annotation, incorrect size, and position of bounding box. These kinds of annotation errors have been found for many images in MS-COCO as well as in Open Image Dataset.

In this paper, we present a method to enhance the quality of ground truth annotations for human object detectors. Our approach aims to address annotation errors in both MS-COCO and Open Image Dataset, as depicted in Fig. 1. The improvement of ground truth annotations was achieved through the following process. Firstly, we removed annotations of grouped objects. Then, we added missing annotations to individual instances caused by annotators and removed group objects. Consequently, we utilized multiple object detection models to generate bounding boxes, comparing them with the ground truth bounding boxes to calculate the confidence score for each bounding box. The bounding box with the highest confidence score is selected to be a new ground truth annotation. We define the confidence score based on our own perspective during the calculation. In this study, we employed the You Only Learn One Representation (YOLOR) [12] as one of the human object detectors. YOLOR achieved a comparable performance to Scaled-YOLOv4-P7 [13], while also improving inference speed by 88%. Additionally, we incorporated Faster-RCNN in our study, which merged with a Region Proposal Network (RPN) by sharing their convolutional features. Faster-RCNN has demonstrated state-of-the-art object detection accuracy on PASCAL VOC 2007, 2012, and MS-COCO datasets [14]. Furthermore, we implemented EfficientDet [4] and RetinaNet [15] in our study to have diverse models for evaluation.

In summary, our proposed technique for improving ground truth annotations provides the following contributions:

- We have developed an automatic method to generate a better quality of ground truth annotations based on machine learning models.
- We solve annotation issues in public datasets, particularly for the person class, which can be utilized to address various problems such as human detection, recognition, tracking, and other types of analysis. Furthermore, our approaches are applicable in other classes of object detection datasets, dealing with single-class tasks.
- We offer a solution for researchers who adopt a model-centric approach, allowing them to access a large number of improved quality datasets for their model training.

## 2 Related works

Data-centric AI is a newly prominent term among researchers in artificial intelligence systems. Over the past few decades, much of the research has been concentrated on conventional model-centric approaches aimed at enhancing the performance of machine learning models. These approaches have led to the development of numerous models with modern network architectures. The next phase in this journey involves a shift towards a data-centric approach, which focuses on improving data quality. This approach is key to unlocking the potential of the new generation of AI systems [1].

In recent times, numerous studies have been exploring a novel approach aimed at improving dataset quality to enhance model performance. Dataset quality improvement can be achieved through various methods, including manual re-annotation of ground truth with improved guidelines, model adaptation to identify and filter out noisy data, and the use of machine learning models to generate higher-quality annotations. Ma et al. [5] conducted a study to manually re-annotate MS-COCO and Open Image Dataset according to their specific guidelines, which resulted in an improvement in the quality of these public datasets. Their new annotations for Open Image Dataset led to enhanced model performance, as demonstrated through experimentation with five detectors (Faster-RCNN [14], SSD [16], YOLOv3 [17], EfficientDet [4], and DETR [18]). However, the study did not indicate similar improvements for MS-COCO. It is worth noting that the manual re-annotation process required a significant amount of time. Furthermore, the study highlighted that annotation remains a major issue for these public datasets. Additionally, machine learning models can be employed to identify the dataset issues instead of relying solely on manual efforts.

Moreover, other studies have attempted to address data annotation challenges by focusing on model adjustments to learn from datasets containing errors. For example, "Confident Learning" focused on label quality by characterizing and identifying label errors in datasets [19]. This approach utilized confident joint to estimate the joint distribution between noisy given labels and unknown labels, achieving performance that exceeded the seven recent competitive approaches in learning with noisy labels on CIFAR datasets [20].

Another approach involves a hybrid supervised learning framework [21] designed to tackle missing label problems. This framework introduces weakly supervised object detection (WSOD) as a teacher model to generate pseudo labels, which are then merged with ground truth to train a new object detector. The hybrid framework iteratively replaces the teacher model with a newly trained object detector and updates pseudo labels for the next training iteration. Experimental results on Pascal VOC 2007 and 2012 datasets demonstrated the method's superior performance in different levels of missing instances.

Furthermore, another study [11] proposed a solution for training object detectors using datasets with noisy annotations. This framework, based on Faster R-CNN, reduces noise by correcting annotations during the training process. Center-Matching Correction was used to calculate the similarity between ground truth bounding boxes and region proposals from Faster R-CNN. Additionally, cross-iteration noise judgment was applied to distinguish between correct and incorrect class labels.

While many studies have demonstrated significant improvements in object detection models by improving dataset quality through data-centric approaches, there are still limitations. Manual annotation remains time-consuming, and the results are still unsatisfactory. Furthermore, automated solutions using models for annotation error correction often focus on enhancing their own model's performance without rectifying the original dataset's inaccurate annotations.

Taking into account the issues outlined, we propose an approach to automatically address annotation errors in the public datasets by directly updating inaccurate annotations. Our method aims to address missing annotations, resize incorrectly sized bounding boxes, and relocate inaccurately placed bounding boxes.

## 3 Methods

In solving the issue of missing and incorrect bounding boxes in public datasets, there is another concern that can affect model performance. This concern arises from annotation practices, where annotators attempt to group crowded instances within a single bounding box and assign a single label to them as shown in Fig. 2. It illustrated most areas of bounding box cover background and other instances that belong to different classes. This makes the model learn with the wrong information of the target class and increases the number of false positive. This can lead to inaccuracies in the training and validating models, which can negatively affect the ability to detect and recognize individual objects. So far, most researchers have primarily focused on improving the model without paying significant attention to the annotation issue. They generally download public datasets and keep the annotations fixed as the original ground truth to train and evaluate their models. In the annotation files of MS-COCO and Open Image Dataset, there is an attribute labeled as “isCrowded", which indicates that the bounding box covers multiple instances. However, removing annotations of grouped instances from the datasets can help improve the performance of the model. Based on the results of our experiments shown in Tables 2 and 3, it is evident the performance of each model improved after removing only the annotations of grouped instances. All models were used without updating their structures or modifying their parameters and were trained for 50 epochs using the datasets listed in Table 1. Moreover, the models trained with Open Image Dataset showed significantly improvement compared to those trained with MS-COCO Dataset when annotations of grouped instances were removed, as shown in Tables 2 and

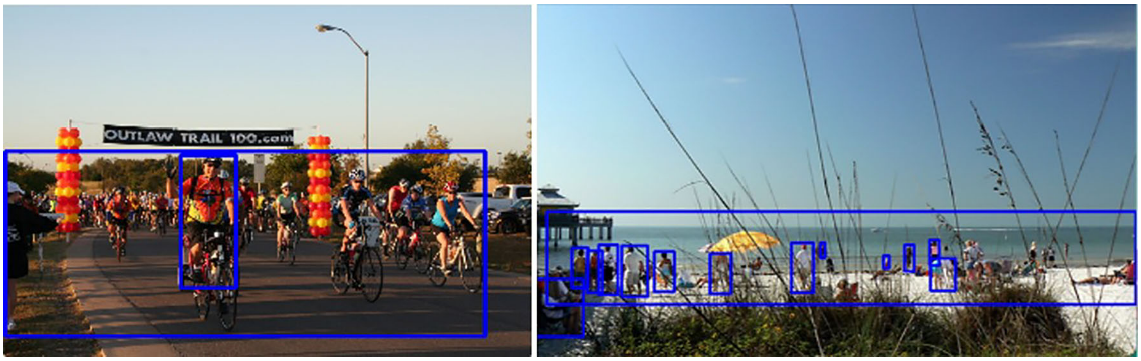


**Fig. 2** Grouped instances of bounding boxes. It shows that annotators try to group many humans' instances in the crowded as one bounding box and label. It illustrated most area of bounding box cover background and other instances that are belong to different classes. More importantly, this made model to learn with the wrong information of the target class and false positive will be increased which results the lower mAP

**Table 1** Public datasets information. The number of images with the amounts of instances in parenthesis

| | MS-COCO | | Open Image | |
|---|---|---|---|---|
| | Original* | Removed** | Original* | Removed** |
| Train | 25,000 (102,235) | 25,000 (100,217) | 25,000 (103,884) | 24,291 (97,988) |
| Validation | 2693 (11,004) | 2693 (10,777) | 5000 (10,360) | 3984 (8604) |

* the dataset with grouped instances ** the dataset after grouped instances are removed

**Table 2** Model performance comparison between original datasets and grouped instances annotation elimination of MS-COCO

| | MS-COCO | | | |
|---|---|---|---|---|
| Models | AP0.5[a] | | mAP[b] | |
| | Original | Removed | Original | Removed |
| YOLOR | 0.845 | **0.852** | 0.605 | **0.615** |
| Faster-RCNN | 0.737 | **0.752** | 0.450 | **0.458** |
| EfficientDet | 0.786 | **0.794** | 0.514 | **0.520** |
| RetinaNet | 0.649 | **0.659** | 0.368 | **0.375** |

[a] The average precision with Intersection over Union (IoU) 0.5

[b] The mean average precision with IoU[0.5:0.5:0.95], start from 0.5 to 0.95 with the step 0.5

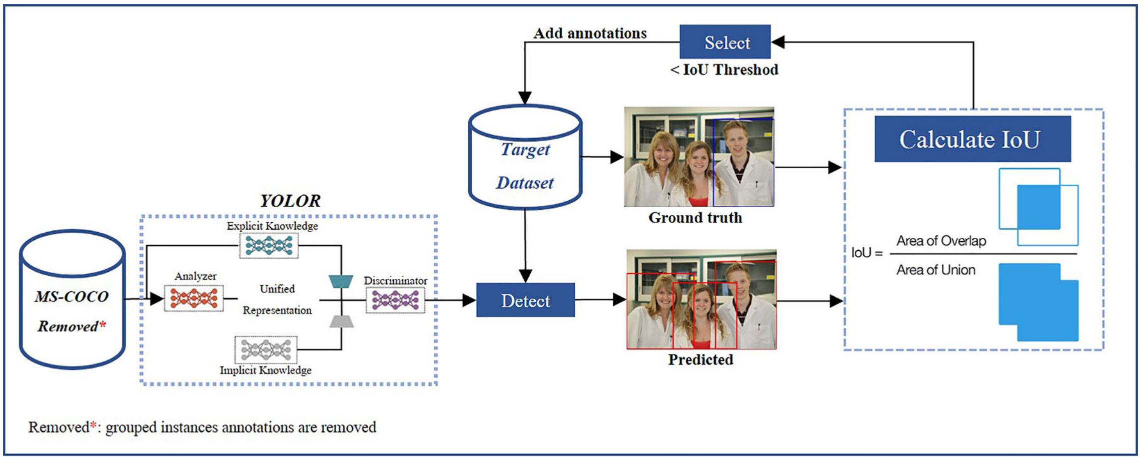
**Table 3** Model performance comparison between original datasets and grouped instances annotation elimination of Open Image Dataset

| | Open Image | | | |
|---|---|---|---|---|
| Models | AP0.5 | | mAP | |
| | Original | Removed | Original | Removed |
| YOLOR | 0.466 | **0.654** | 0.337 | **0.477** |
| Faster-RCNN | 0.457 | **0.613** | 0.269 | **0.368** |
| EfficientDet | 0.496 | **0.650** | 0.319 | **0.423** |
| RetinaNet | 0.435 | **0.560** | 0.234 | **0.306** |

3. This is because there were more grouped instances in the former than in the latter, as indicated in Table 1. These findings highlight the effectiveness of removing inappropriate annotations. However, other types of erroneous annotations mentioned in Fig. 1 still exist and need to be solved for improving the data quality which is our goal in this study.

### 3.1 Adding missing annotation approach

After the annotations of grouped instances are removed, those instances are put in a status with no bounding boxes and class labels, as shown in Fig. 1a. This deteriorates the quality of the dataset by increasing the number of missing annotations. To address this problem, we proposed a technique, as illustrated in Fig. 3, that utilizes a trained object detector to add missing annotations of the instances. Firstly, we selected a model trained by YOLOR with the MS-COCO dataset, in which annotations of grouped instances were removed. The reason for this selection was that this model provided the best result based on mAP when compared to other models and datasets that we used, as shown in Tables 2 and 3. Next, we executed the trained model to detect instances of individual human objects in the dataset and generated bounding boxes for them. Subsequently, we compared each of these bounding boxes with the original ground truth bounding boxes. The comparison was performed by calculating the Intersection over Union (IoU) [22], which measures the similarity between the two bounding boxes. We set a threshold value of IoU to decide whether the bounding boxes were the same or not. Figure 4 shows the similarity between two bounding boxes based on IoU. If a bounding box had an IoU smaller than the threshold with any other bounding box existing in the original dataset, it was assumed to be a new one and



Removed*: grouped instances annotations are removed

**Fig. 3** Proposed technique to add missing annotations. YOLOR model is trained with MS-COCO dataset then the model is used to detect the annotations in a target dataset. Next, IoU between detected and ground truth bounding boxes is calculated. Finally, the select is made by comparing the IoU with the threshold value to determine whether the bounding box is missed one or not. YOLOR's structure is cited from [12]

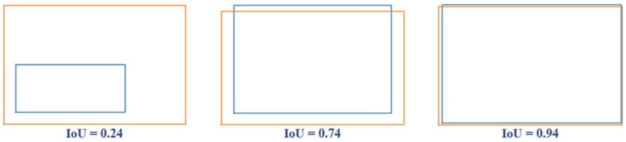


**Fig. 4** Example of similarity between two bounding boxes based on IoU (Intersection over Union)

added to the dataset. By applying this technique, the dataset was retrieved with better quality of annotations, which consequently enhanced the performance of the model.

## 3.2 Correcting size and position of bounding box approach

Although missing annotations have been fairly addressed in the previous process, some annotations still contain errors with size and position of bounding boxes in the public datasets, as shown in Fig. 1b, c. This is because many annotators with different levels of experiences contributed to annotating the datasets. To address this issue, we propose a new approach and demonstrate its effectiveness using the MS-COCO and Open Image Dataset.

In our method, several machine learning models are used to generate bounding boxes. Afterwards, the similarity of each pair of all generated bounding boxes including the ground truth is calculated based on IoU. As mentioned previously, the MS-COCO dataset without grouped instances annotation is the best dataset available, and it is used to train the models in our method. To select a better bounding box for a new alternative ground truth annotation. To simplify the explanation, we first consider the case of using two models shown in Fig. 5. For instance, there are many objects in an image, so it's necessary to match the bounding boxes that belong to the same object for comparison. After finding the bounding boxes generated by both models that correspond to the ground truth, we calculate the IoU using Eq. 1, i.e. $IoU(b_{GT}, b_{M1})$, $IoU(b_{GT}, b_{M2})$, and $IoU(b_{M1}, b_{M2})$. Here, $b_{M1}, b_{M2}, b_{GT}$ are bounding boxes generated from model 1, model 2 and the ground truth, respectively. Next, we calculated the confidence score for each bounding box defined by Eq. 2, where $S_{M1}$, $S_{M2}$, $S_{GT}$ represent the scores of bounding boxes generated by model 1, model 2, and the ground truth, respectively. The confidence score is normalized to range from 0 to 1 which represented the percentage of two boxes overlapping. Finally, we make an enhanced ground truth bounding box by selecting the one with the highest-scoring using Eq. 3 as a new alternative ground truth annotations.

The generalization of Eq. 2, when $n$ models are used, is denoted as Eq. 4. Here the set $D = \{b_{M1}, b_{M2}, b_{M3}, \ldots, b_{Mn}, b_{GT}\}$ is a set of bounding boxes of an object detected by $n$ models and the ground truth. The confidence score $S_{b_m}$ for a bounding box $b_m$ is calculated as Eq. 4. The score is normalized, ranging from 0 to 1, by the normalization factor $\alpha$.

$$IoU(b_1, b_2) = \frac{b_1 \cap b_2}{b_1 \cup b_2} \tag{1}$$

Definition of the confidence score S for a bounding box generated by each detection model.

$$\begin{aligned} S_{M1} &= \frac{1}{2}(IoU(b_{M1}, b_{GT}) + IoU(b_{M1}, b_{M2})) \\ S_{M2} &= \frac{1}{2}(IoU(b_{M2}, b_{GT}) + IoU(b_{M2}, b_{M1})) \\ S_{GT} &= \frac{1}{2}(IoU(b_{GT}, b_{M1}) + IoU(b_{GT}, b_{M2})) \end{aligned} \tag{2}$$

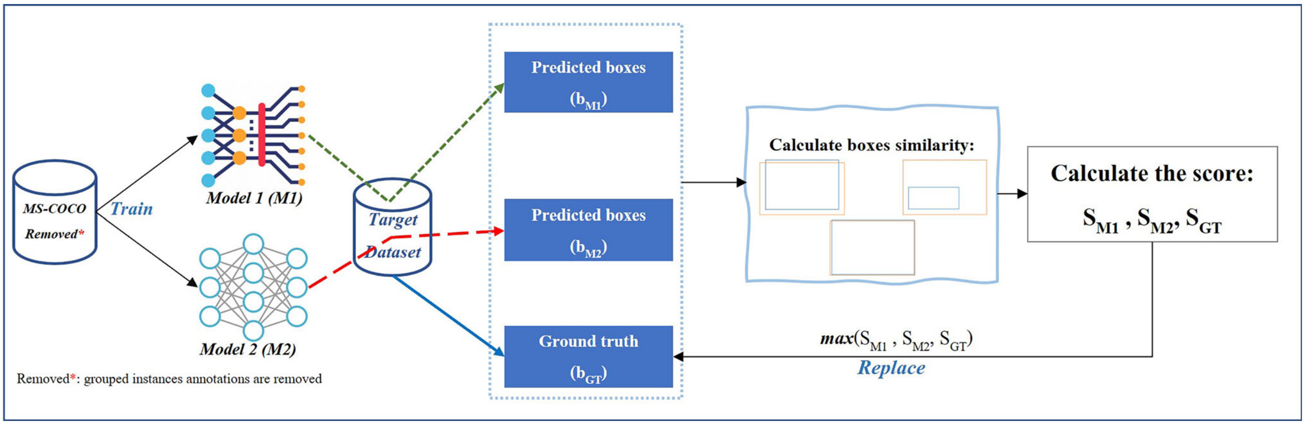

MS-COCO
Removed*
Train
Model 1 (M1)
Model 2 (M2)
Target Dataset
Predicted boxes ($b_{M1}$)
Predicted boxes ($b_{M2}$)
Ground truth ($b_{GT}$)
Calculate boxes similarity:
Calculate the score:
$S_{M1}$, $S_{M2}$, $S_{GT}$
max($S_{M1}$, $S_{M2}$, $S_{GT}$)
Replace
Removed*: grouped instances annotations are removed

**Fig. 5** The proposed structure for correcting errors of the size and position of bounding boxes. $b_{M1}$, $b_{M2}$, and $b_{GT}$ are the bounding boxes predicted from model 1, model 2 and from ground truth, respectively. $S_{M1}$, $S_{M2}$, and $S_{GT}$ are the confidence scores calculated for the bounding boxes of model 1, model 2, and ground truth, respectively

$b_{M1}, b_{M2}, b_{GT}$: bounding boxes generated from Model 1, Model 2, and Ground Truth
$S_{M1}, S_{M2}, S_{GT}$: the confidence scores of bounding boxes generated by Model 1, Model 2, and Ground truth

$$new_{b_{GT}} = \{box(S) | S = max(S_{M1}, S_{M2}, S_{GT})\} \qquad (3)$$

$box(S)$ is a box corresponding with score $S$

$$S_{b_m} = \alpha \sum_{b_i \neq b_m} IoU(b_m, b_i) \qquad (4)$$

$where \quad b_m, b_i \in D; D = \{b_{M1}, b_{M2}, b_{M3}, \ldots, b_{Mn}, b_{GT}\}$
$\alpha = \frac{1}{C_2^{n+1}} \quad where \quad n+1$ is the number of bounding boxes generated from n models plus ground truth.

Then, Eq. 3 is generalized to Eq. 5.

$$new_{b_{GT}} = \{box(S) | S = max(S_{M1}, S_{M2}, \ldots, S_{Mn}, S_{GT})\} \qquad (5)$$

## 4 Results and discussion

In this section, we describe the experimental results and discussion. The experiment of training and validation of models was performed by the machine with the environment below.

- OS: Ubuntu 18.04.5 LTS
- CPU: Intel Core i9-10980XE(3.0GHz)
- Memory: 128GB, 2 GPU: Geforce RTX A6000
- CUDA: 11.3, Python: 3.10.4, PyTorch: 1.11.0

### 4.1 Adding missing annotation

For our proposed approach to add missing annotations in the ground truth datasets, we conducted experiments on MS-COCO and Open Image Dataset using YOLOR model (shown in Fig. 3). In the validation process, we used other three models such as Faster R-CNN, EfficientDet, and RetinaNet to check its performance after the datasets are improved. As mentioned in the previous section, a threshold value of IoU between two boxes is introduced to determine whether the two objects are the same or not. If a bounding box has an IoU smaller than the threshold with any other bounding box existing in the same image of the original dataset, it is assumed to be a new bounding box. Therefore, the bounding box is considered appropriate to add to the dataset as the ground truth. In the experiment, we set the threshold values of 0.5 and 0.75.

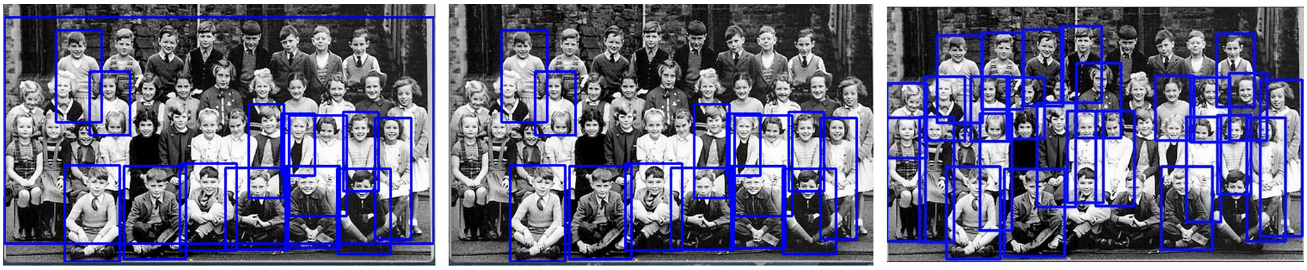
As the results of validation, Table 4 shows the mAP of the three models we used to figure out the quality of MS-COCO which is generated by our approach. The results illustrate that our proposed approach which adds missing annotation to the ground truth is outperform in those three models compared to the original dataset. It also indicates that threshold 0.5 gives better performance compared to threshold 0.75 in MS-COCO dataset for both AP0.5 and mAP evaluation. Figure 6 displays the results after we applied our approach to add missing annotations compared to the original annotations. It demonstrates significant improvement in the quality of the new annotations.

We used the same three models as MS-COCO to validate the performance of our improved dataset on the Open Image Dataset. The results in Table 5 indicate that our proposed approach to add missing annotation in Open Image Dataset also outperforms for those three models. Addition-

**Table 4** The evaluation of MS-COCO after we added the missing annotation to ground truth. The result shows the comparison among three models with threshold 0.5 and 0.75

| | MS-COCO | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Models | AP0.5 | | | | mAP | | | |
| | Ori | Remo | Add missing (Our) | | Ori | Remo | Add missing (Our) | |
| | | | Thr0.5 | Thr0.75 | | | Thr0.5 | Thr0.75 |
| Faster-RCNN | 0.738 | 0.753 | **0.783(↑5%)** | 0.7366 | 0.451 | 0.459 | **0.468(↑2%)** | 0.458 |
| EfficientDet | 0.786 | 0.794 | **0.817(↑3%)** | 0.776 | 0.514 | 0.520 | **0.525(↑1%)** | 0.510 |
| RetinaNet | 0.649 | 0.659 | **0.664(↑2%)** | 0.633 | 0.368 | 0.370 | **0.370(↑0.2%)** | 0.365 |

Ori: original dataset with grouped instances
Remo: the dataset after grouped instances are removed
Thr0.5, Thr0.75: threshold value set 0.5 and 0.75, respectively
(↑%) means the performance increases by adding missing bounding boxes when compared with the original dataset



(a) Original

(b) After removing grouped instance boxes

(c) Our proposed adding missing boxes

**Fig. 6** The results of our proposed adding missing bounding boxes to MS-COCO. **a** Is the original bounding boxes from the dataset. **b** Shows the bounding boxes of every single instance remaining after removing the bounding box which covered many instances. It shows many missing bounding boxes which means that many instances haven't been annotated. **c** Displays the results of our proposed approach to add missing annotation. We can see that almost of the instances are annotated

ally, we found that threshold 0.5 gives the best AP0.5 while threshold 0.75 gives the best mAP for all three models. Our proposed approach increased 36%, 35% and 27% for Faster-RCNN, EfficientDet and RetinaNet, respectively in term of AP0.5 compared to the original dataset. Furthermore, the mAP increased by 22%, 24% and 17% for Faster-RCNN, EfficientDet and RetinaNet, respectively compared to the original dataset. Figure 7 shows the missing bounding boxes that our proposed approach added to the original Open Image dataset, indicating a significant improvement in model performance with a better-quality dataset.

## 4.2 Correcting size and position of bounding boxes

In the previous section, we presented the effectiveness of adding missing bounding boxes to MS-COCO and Open Image Dataset. There are still issues with incorrect size and position of bounding boxes in these datasets (shown in Fig. 1b, c). In this section, we will illustrate how to solve these issues with our proposed approach (shown in Fig. 5) using Eqs. from 1 to 3.

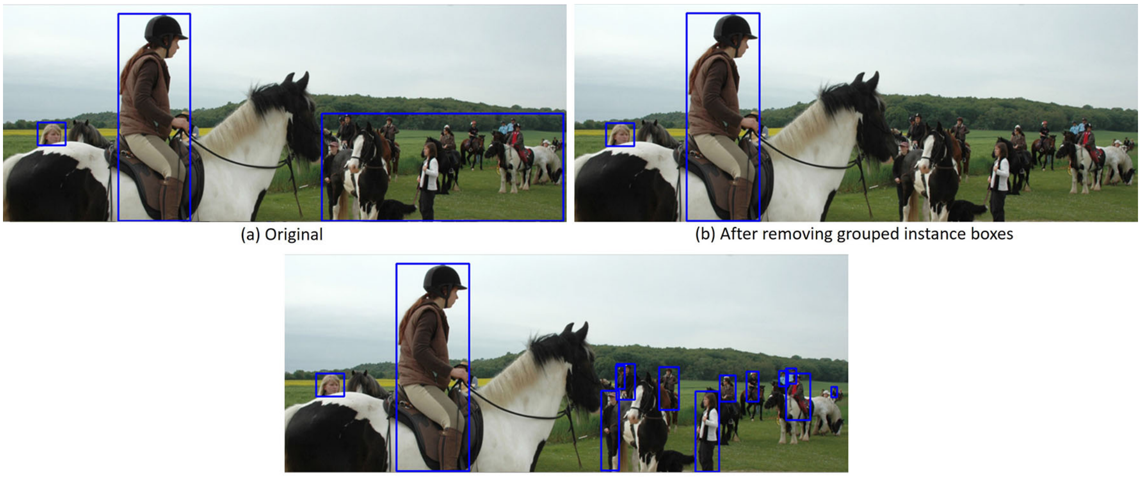
We used two trained models to calculate the confidence scores of their predicted bounding boxes and the ground truth to which missing boxes were added in the previous section. Subsequently, we selected the bounding box with the highest confidence score to be new ground truth. In order to check the effectiveness of this approach, we performed two experiments. The first experiment involved using the pair of YOLOR and Faster-RCNN, while the second experiment utilized the pair of YOLOR and EfficientDet to generate an improved ground truth. Our approach aimed to enhance the ground truth by correcting the size and position of bounding boxes. After that, we compared these two experiments to evaluate the effect of our method for enhancing ground truth annotations.

### 4.2.1 Results with MS-COCO

As shown in Table 6, we compared the results of our new approach with the original dataset and our previous method which only added the missing bounding boxes. In the previous approach, we achieved the best result (in terms of mAP) by setting the threshold value 0.5 in MS-COCO for all three models, compared to the original dataset. Therefore, we also set the threshold to 0.5 and compared the mAP for this approach and the previous method. The experimental results indicate mAPs are increased compared to the cases of only adding missing bounding boxes, which implies that a better data quality is achieved by our approach. And it even better if we compared to the original dataset.

**Table 5** The evaluation of Open Image Dataset after adding the missing annotation to ground truth. It is compared between three models with threshold 0.5 and 0.75

| | Open Image | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Models | AP0.5 | | | | mAP | | | |
| | Ori | Remo | Add missing (Our) | | Ori | Remo | Add missing (Our) | |
| | | | Thr0.5 | Thr0.75 | | | Thr0.5 | Thr0.75 |
| Faster-RCNN | 0.458 | 0.614 | **0.819**(↑**36**%) | 0.777 | 0.270 | 0.369 | 0.490 | **0.490**(↑**22**%) |
| EfficientDet | 0.496 | 0.650 | **0.845**(↑**35**%) | 0.821 | 0.319 | 0.423 | 0.547 | **0.559**(↑**24**%) |
| RetinaNet | 0.435 | 0.560 | **0.710**(↑**27**%) | 0.699 | 0.234 | 0.306 | 0.380 | **0.399**(↑**17**%) |


(a) Original

(b) After removing grouped instance boxes

(c) Our proposed adding missing boxes

**Fig. 7** The results of our proposed adding missing bounding boxes to Open Image Dataset. **a** Is the original bounding boxes from the dataset. **b** shows the bounding boxes of every single instance remaining after removing the bounding box which covered many instances. It shows many missing bounding boxes which means that many instances haven't been annotated. **c** Displays the results of our proposed approach to add missing annotation. We can see that almost of the instances are annotated

**Table 6** The evaluation of the effect by correcting size and position of bounding boxes with MS-COCO dataset

| | MS-COCO | | | |
|---|---|---|---|---|
| Models | mAP | | | |
| | Original | Add missing (Our) | Correcting bounding boxes (Our) | |
| | | | YOLOR+Faster-RCNN | YOLOR+EfficientDet |
| Faster-RCNN | 0.450 | 0.468 | **0.536** (↑**7**%)(↑↑**9**%) | 0.528 (↑6%)(↑↑8%) |
| EfficientDet | 0.514 | 0.525 | 0.584 (↑6%)(↑↑7%) | **0.617** (↑**9**%)(↑↑**10**%) |
| RetinaNet | 0.368 | 0.370 | **0.438** (↑**7**%)(↑↑**7**%) | 0.429 (↑6%)(↑↑6%) |

(↑%) means the performance increases by correcting bounding boxes when compared with our method for adding missing. (↑↑%) means the performance increases by correcting bounding boxes when compared with the original dataset

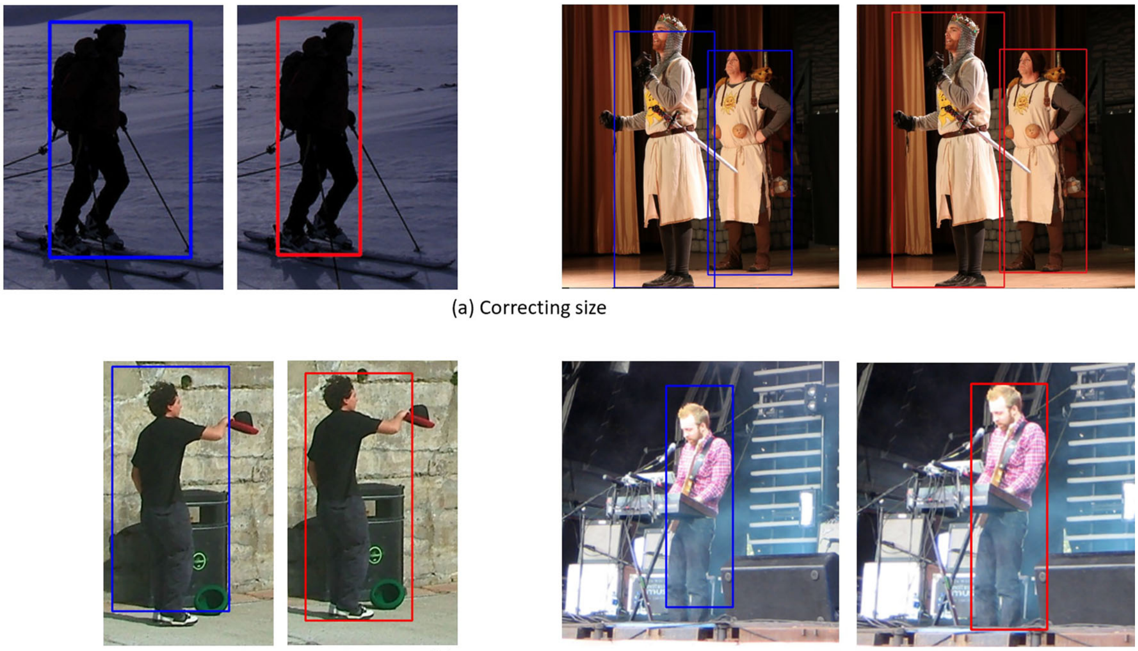
#### 4.2.2 Results with Open Image Dataset

Table 7 shows the comparison of the result of our new approach of correcting size and position bounding boxes with Open Image Dataset. Because the adding missing bounding boxes approach achieved the best results with threshold value 0.75 in Open Image Dataset, we compared only this result with our new method. As the case with MS-COCO, the experimental results indicate mAPs are increased compared to the case of only adding missing bounding boxes, which means that a better data quality is achieved by our approach not only with MS-COCO but also with Open Image Dataset. Surprisingly, it shows a big improvement if we compared to original dataset.

In Fig. 8 illustrates the results of our proposed methods to correct the size and position of bounding boxes for both MS-COCO and Open Image Dataset. Figure 8a shows the incorrect bounding boxes size of the objects which are displayed in blue boxes and those boxes are improved by our approach. The red bounding boxes are fitter to objects compared to the original. Figure 8b conveys the correct position of red bounding boxes which are generated by our approach while blue bounding boxes are from original dataset.

**Table 7** The evaluation of the effect by correcting size and position of bounding boxes with Open Image Dataset

| | Open Image | | | |
|---|---|---|---|---|
| Models | mAP | | | |
| | Original | Add missing (Our) | Correcting bounding boxes (Our) | |
| | | | YOLOR+Faster-RCNN | YOLOR+EfficientDet |
| Faster-RCNN | 0.269 | 0.490 | **0.554** (↑**6**%)(↑↑**29**%) | 0.551 (↑6%)(↑↑28%) |
| EfficientDet | 0.319 | 0.559 | 0.604 (↑5%)(↑↑29%) | **0.637** (↑**8**%)(↑↑**32**%) |
| RetinaNet | 0.234 | 0.399 | **0.460** (↑**6**%)(↑↑**23**%) | 0.447 (↑5%)(↑↑21%) |



(a) Correcting size

(b) Correcting position

**Fig. 8** The result of size and position correction. The blue bounding boxes represented original annotations while red boxes are the results of our method by correcting size and position

### 4.3 Ablation studies

To validate the effectiveness of our approaches, we conducted ablation studies using our improved datasets. While the training sets remain unchanged, retaining the original data from MS-COCO and Open Image datasets, our approaches focus only on improving the validation sets in this ablation study. These original training sets are used to train various models, such as Faster-RCNN [14], EfficientDet [4], RetinaNet [15], YOLOv7 [23], DINO [24], and RT-DETR [25] with our configuration setting. Subsequently, these trained models are evaluated using the improved datasets generated by our approaches, and their performance is compared to that of the same models evaluated with the original validation sets from MS-COCO and Open Image datasets.

Moreover, we employed pre-trained models provided by Faster-RCNN [14], EfficientDet [4], RetinaNet [15], YOLOv7 [23], DINO [24], and RT-DETR [25] to evaluate our improved validation sets against the original ones. The results, as presented in Tables 8 and 9, demonstrate the enhancement of all the models, including those pre-trained, when evaluated with our validation sets compared to the original ones. These results underscore the effectiveness of our approaches in improving dataset quality such as MS-COCO and Open Image under the condition of annotation fallacies. More specifically, our approach for correcting bounding box size and location using YOLOR + EfficientDet produces the most favorable annotations among the approaches, as demonstrated by the evaluation results illustrated in Experiment$^2$ of Tables 8 and 9.

### 4.4 Discussion

In terms of datasets quality problems, there are various approaches to solve them depending on the type of data and the specific problem. One of the most accurate ways to address data annotation issues is through manual work, as proposed by Ma et al. [5]. However, this method requires clear guidelines that align with the capabilities of the validation models. Additionally, it is expensive due to the need

**Table 8** The ablation studies on MS-COCO dataset

| | | MS-COCO | | |
|---|---|---|---|---|
| Models | Train | Validation (mAP) | | |
| | | Original†† | Experiment[a] (Our) | Experiment[b] (Our) |
| **Faster-RCNN** | Pre-trained† | 0.446 | 0.476 | **0.482** |
| | Original†† | 0.450 | **0.515** | 0.510 |
| **EfficientDet** | Pre-trained† | 0.555 | 0.609 | **0.625** |
| | Original†† | 0.514 | 0.549 | **0.574** |
| **RetinaNet** | Pre-trained† | 0.457 | 0.489 | **0.497** |
| | Original†† | 0.368 | 0.382 | **0.384** |
| **YOLOv7** | Pre-trained† | 0.592 | 0.670 | **0.677** |
| | Original†† | 0.532 | 0.605 | **0.612** |
| **DINO** | Pre-trained† | 0.581 | 0.653 | **0.657** |
| | Original†† | 0.552 | 0.618 | **0.623** |
| **RT-DETR** | Pre-trained† | 0.601 | 0.660 | **0.667** |
| | Original†† | 0.536 | 0.588 | **0.593** |

†Indicate the pre-trained model provided from the reference paper.

††indicate the original datasets (shown in Table 1) are used to train/validate the model.

[a] indicate that our improved validation set (add missing and correct error using YOLOR+Faster-RCNN) is validated.

[b] indicate that our improved validation set (add missing and correct error using YOLOR+EfficientDet) is validated

for highly trained annotators, and it is time-consuming as mentioned in the study [5].

If we compare our study with the study of Ma et al. [5], we spent only around 2 h (excluding model training) to enhance annotations by automatically adding missing bounding boxes, correcting their size and position for 27k images from the MS-COCO dataset using our mid-range desktop computer with the environment previously mentioned. In contrast, they spent 19 days manually re-annotating 80k images from MS-COCO based on their guidelines by employing hundreds of human annotators. Their approach required significant financial resources and time, yet their new annotations for MS-COCO did not lead to improved model performance. On the other hand, our proposed approach demonstrated improved model performance in MS-COCO, as shown in Tables 4 and 6.

Similarly, when working with the Open Image Dataset, Ma et al. [5] spent 9 days involving 40 annotators and quality inspectors to annotate just 5k images. In contrast, we spent approximately 2 h (excluding model training) to process 29k images using our proposed method on a mid-range desktop computer. In the experiments, our proposed method outperformed compared to their manual work based on mean Average Precision (mAP) in the person class. Moreover, our proposed methods required significantly less time and resources compared to their study while achieving better results in both MS-COCO and Open Image Dataset.

From the result of bounding boxes size and position correction, Tables 6 and 7 demonstrate that our approach to correcting the position and size of bounding boxes achieve up to 9% improvement for MS-COCO while Open image Dataset can achieve up to 8% compared to our methods which is just adding the missing annotation. Notably, we reached 14% in MS-COCO and 43% in Open Image Dataset compared to the original datasets. This indicates that size and position errors happen similarly in both datasets while missing annotations occur in Open Image Dataset much more than MS-COCO (shown in Tables 4 and 5).

However, there is still room for improvement in our proposed methods, particularly in the aspect of adding missing annotations. Currently, we added new annotations to the ground truth solely based on the model we trained. It would be advantageous to incorporate information from the grouped instances, which we excluded prior to applying this proposed approach. This additional information can be used in conjunction with the model to determine whether to add new bounding boxes to the ground truth.

Furthermore, another annotation issue in the MS-COCO dataset that requires manual handling and removal is the annotation of excessively small human body parts, as depicted in Fig. 9. This significantly impacts model performance, as also mentioned in the study [5]. Similarly, in the Open Image Dataset, there are instances of unreal human objects (as shown in Fig. 10) that exist in the dataset, leading to learning errors in the model. Therefore, addressing these specific problems present in both datasets would enhance the data quality, ultimately improving model performance within

**Table 9** The ablation studies on Open Image dataset

| | | Open Image | | |
|---|---|---|---|---|
| Models | Train | Validation (mAP) | | |
| | | Original†† | Experiment[a] (Our) | Experiment[b] (Our) |
| **Faster-RCNN** | Pre-trained† | 0.220 | 0.483 | **0.486** |
| | Original†† | 0.270 | **0.466** | 0.461 |
| **EfficientDet** | Pre-trained† | 0.293 | 0.625 | **0.634** |
| | Original†† | 0.320 | 0.486 | **0.493** |
| **RetinaNet** | Pre-trained† | 0.250 | **0.508** | **0.508** |
| | Original†† | 0.234 | 0.327 | **0.329** |
| **YOLOv7** | Pre-trained† | 0.276 | 0.694 | **0.698** |
| | Original†† | 0.283 | 0.556 | **0.557** |
| **DINO** | Pre-trained† | 0.278 | 0.661 | **0.664** |
| | Original†† | 0.398 | 0.595 | **0.596** |
| **RT-DETR** | Pre-trained† | 0.315 | 0.665 | **0.672** |
| | Original†† | 0.376 | 0.545 | **0.548** |

†Indicate the pre-trained model provided from the reference paper

††indicate the original datasets (shown in 1) are used to train/validate the model

[a]indicate that our improved validation set (add missing and correct error using YOLOR+Faster-RCNN) is validated

[b] indicate that our improved validation set (add missing and correct error using YOLOR+EfficientDet) is validated

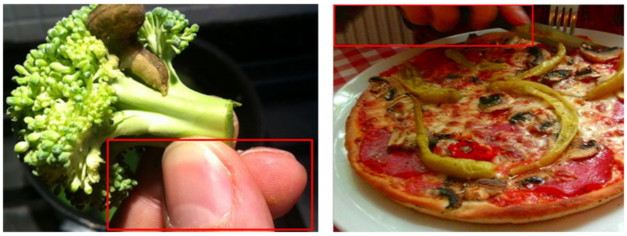


**Fig. 9** Annotation of too small human's part in MS-COCO

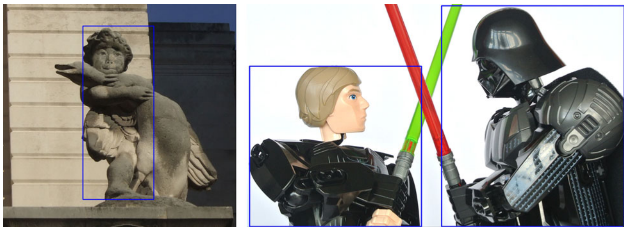


**Fig. 10** Annotation of unreal human in Open Image Dataset

the data-centric approach. These aspects will be considered in our future work.

## 5 Conclusions

In this study, our proposed approach aims to enhance the data quality in both the MS-COCO and Open Image Dataset by addressing missing annotations as well as errors in the size and position of bounding boxes. Additionally, we have eliminated grouped annotations from the original datasets, as they significantly impact the performance of the models, as demonstrated in the experiments. However, our approach for adding missing annotations to the ground truth still encounters some issues, such as redundant annotations where a single object has multiple annotations, although this is a rare occurrence. While this does not heavily affect the model's performance, addressing this problem in the future would lead to even better model performance. Nevertheless, our proposed methods, which involve improving the size and position of bounding boxes, achieve higher accuracy at a lower cost compared to the manual work proposed by other studies.

**Author Contributions** Conceptualization and Methodology: Sotheany Nou; Writing - original draft preparation: Sotheany Nou; Writing - review and editing: Sotheany Nou, Joong-Sun Lee; Commentator on manuscript: Nagaaki Ohyama; Guidance and supervision:Takashi Obi. All authors approved the final manuscript to submit.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Databricks: Data centric AI development from big data to good data Andrew NG data+AI summit (2022). https://www.youtube.com/watch?v=avoijDORAlc
2. Qiao, S., Chen, L., Yuille, A.: Detectors: detecting objects with recursive feature pyramid and switchable atrous convolution (2021)
3. Long, X., et al.: Pp–yolo: an effective and efficient implementation of object detector. CoRR (2020). arXiv:abs/2007.12099
4. Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection (2020)
5. Ma, J., Ushiku, Y., Sagara, M.: The effect of improving annotation quality on object detection datasets: a preliminary study (2022)
6. Lin, T.-Y., et al.: Microsoft coco: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision—ECCV 2014, pp. 740–755. Springer, Cham (2014)
7. Everingham, M., et al.: The Pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**, 98–136 (2015)
8. Deng, J., et al.: ImageNet: a large-scale hierarchical image database (2009)
9. Kuznetsova, A., et al.: The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. Int. J. Comput. Vis. **128**, 1956–1981 (2020)
10. Enderes, S.: The impact of annotation errors on neural networks (2021). https://understand.ai/blog/annotation/machine-learning/autonomous-driving/2021/06/01/impact-of-annotation-errors-on-neural-networks.html
11. Mao, J., Yu, Q., Yamakata, Y., Aizawa, K.: Noisy annotation refinement for object detection. In: Paper Presented at the Conference of the 32nd British Machine Vision Conference (2021)
12. Wang, C.-Y., Yeh, I.-H., Liao, H.: You only learn one representation: unified network for multiple tasks. J. Inf. Sci. Eng. **39**, 691–709 (2021)
13. Wang, C., Bochkovskiy, A., Liao, H.: Scaled-yolov4: scaling cross stage partial network (2021)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2017)
15. Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection (2017)
16. Liu, W., et al. SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision—ECCV 2016, pp. 21–37. Springer, Berlin (2016)
17. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. CoRR (2018). arXiv:abs/1804.02767
18. Carion, N., et al.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision—ECCV 2020, pp. 213–229. Springer, Cham (2020)
19. Northcutt, C.G., Jiang, L., Chuang, I.L.: Confident learning: estimating uncertainty in dataset labels. J. Artif. Intell. Res. (JAIR) **70**, 1373–1411 (2021)
20. Krizhevsky, A.: Learning multiple layers of features from tiny images. Master's Thesis, Department of Computer Science, University of Toronto (2009)
21. Xu, M., Bai, Y., Ghanem, B.: Missing labels in object detection. In: Paper Presented at the Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
22. Rosebrock, A.: Intersection over union (IOU) for object detection (2016). https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/
23. Wang, C., Bochkovskiy, A., Liao, H.: Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors (2023)
24. Zhang, H., et al.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. Comput. Vis. Pattern Recognit. (2022). arXiv:abs/2203.03605
25. Lv, W., et al.: Detrs beat yolos on real-time object detection. Comput. Vis. Pattern Recognit. (2023). arXiv:abs/2304.08069

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Sotheany Nou** received the B.S. degree in information and communication engineering from Institute of Technology of Cambodia, Phnom Penh, in 2015 and M.S. degrees in computing in engineering system from King Mongkut's Institute of Technology Ladkrabang, Bangkok, in 2017. From 2018 to 2021, he was a lecturer at Institute of Technology of Cambodia. He is currently pursuing his doctoral degree in information and communication engineering at Tokyo Institute of Technology, Tokyo. His research interests include computer vision.

**Joong-Sun Lee** received B.S. and M.S. degrees of physics from Yonsei University, the Republic of Korea, in 1984 and 1986 respectively. He received a Ph.D. degree in information processing from Tokyo Institute of Technology in 1995. He had been a senior researcher responsible for developing computer peripherals in the laboratory of LG Electronics. He served as the vice president and head of the research center at an IT vendor in Korea. He also worked for Hitachi and NTT Communications in charge of developing healthcare information systems and smartcard services in Japan. He joined Tokyo Institute of Technology as a specially appointed associate professor in 2008. His research interests include image processing, healthcare information, blockchain applications, and artificial intelligence.

**Nagaaki Ohyama** obtained his Ph.D. from the Department of Information Processing, Tokyo Institute of Technology in 1982 after finishing his B.S. and M.Eng. from the same university in 1977 and 1979, respectively. He is now an institute professor at Institute of Innovative Research of Tokyo Tech. His research areas are information processing, image processing, smart IC card systems, and information systems. His main social responsibilities include activities as a former member of IT Strategic Headquarter. He is the acting chair of the technical committee of the national pension system and security, and responsible for the technical aspects of the national eID card program. He was also a member of Advisory Board on the promotion for utilization of My Number Card and Japanese Public Key Infrastructures.

**Takashi Obi** received his BS degree in physics, his MS and PhD degrees in information physics, from Tokyo Institute of Technology, Japan, in 1990, 1992, and 1997, respectively. He is an Associate Professor of Laboratory for Future Interdisciplinary Research of Science and Technology at the Institute of Innovative Research, Tokyo Institute of Technology. His research focuses on image processing, information systems, medical informatics, and security. He is a member of IEEE, JAMIT, JSAP, JSNM, JSMP and IEICE.