# Drawing the Same Bounding Box Twice? Coping Noisy Annotations in Object Detection with Repeated Labels

David Tschirschwitz[0000−0001−5344−4172],
Christian Benz[0000−0001−9915−0057],
Morris Florek[0009−0008−8425−5161],
Henrik Norderhus[0009−0006−2613−2572],
Benno Stein[0000−0001−9033−2217], and
Volker Rodehorst[0000−0002−4815−0118]

Bauhaus-Universität Weimar, Germany
david.tschirschwitz@uni-weimar.de

**Abstract.** The reliability of supervised machine learning systems depends on the accuracy and availability of ground truth labels. However, the process of human annotation, being prone to error, introduces the potential for noisy labels, which can impede the practicality of these systems. While training with noisy labels is a significant consideration, the reliability of test data is also crucial to ascertain the dependability of the results. A common approach to addressing this issue is repeated labeling, where multiple annotators label the same example, and their labels are combined to provide a better estimate of the true label. In this paper, we propose a novel localization algorithm that adapts well-established ground truth estimation methods for object detection and instance segmentation tasks. The key innovation of our method lies in its ability to transform combined localization and classification tasks into classification-only problems, thus enabling the application of techniques such as Expectation-Maximization (EM) or Majority Voting (MJV). Although our main focus is the aggregation of unique ground truth for test data, our algorithm also shows superior performance during training on the TexBiG dataset, surpassing both noisy label training and label aggregation using Weighted Boxes Fusion (WBF). Our experiments indicate that the benefits of repeated labels emerge under specific dataset and annotation configurations. The key factors appear to be (1) dataset complexity, the (2) annotator consistency, and (3) the given annotation budget constraints.

**Keywords:** Object Detection · Instance Segmentation · Robust Learning.

## 1 Introduction

Data-driven machine learning systems are expected to operate effectively even under "difficult" and unforeseen circumstances. Consider safety-relevant domains

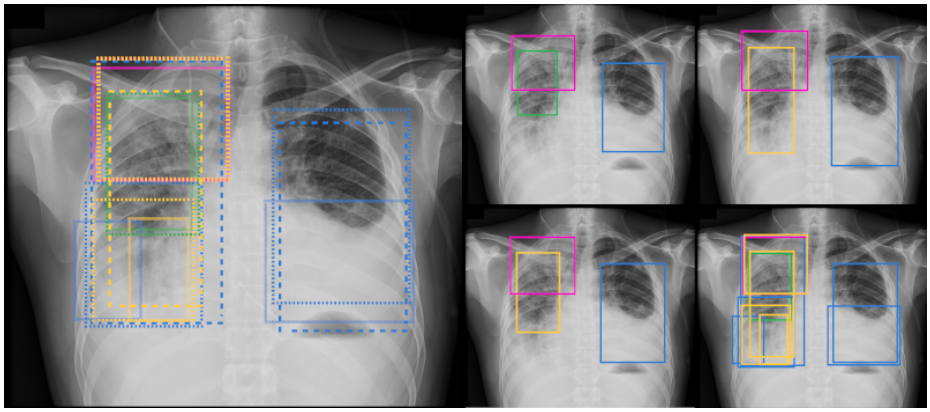

**Fig. 1.** Comparison between different ground truth aggregation methods, exemplary on the VinDr-CXR dataset [23]. Left: the original image with the repeated labels indicated by the different line types. Right: the four smaller images from top left to bottom right are, MJV+$\cap$, LAEM+$\mu$, LAEM+$\cup$ and WBF.

such as autonomous driving, medical diagnosis, or structural health monitoring, where system failure sets lives at risk. Robust systems − those capable of reliable operation in unseen situations − may encounter several challenges, including domain shifts [25,36,16,5], adversarial attacks [43,42], degrading image quality [21,7,41] and noisy or uncertain labels [18,9,10,37]. Past studies [33] indicate that noisy labels can cause more harm than the three aforementioned sources of input noise. Given this context, our study concentrates on addressing the issue of noisy labels, specifically within noisy test data. Without a unique ground truth, evaluation is unattainable. Therefore, to enhance robustness against label noise, it will be pivotal to first devise methods tailored towards annotation aggregation, which lays the groundwork for potential future integration with multi-annotator learning methods.

The creation of annotated data for supervised learning is a costly endeavor, particularly in cases where experts such as medical professionals or domain experts are needed to annotate the data. To mitigate this issue, crowd-sourcing has emerged as a cost-effective means of generating large datasets, albeit with the disadvantage of potentially lower quality annotations that may contain label noise [30,44,47]. Although the reduced costs of crowd-sourced annotations often justifies their use, deep neural networks have the capacity to memorize noisy labels as special cases, leading to a declining performance and overfitting towards the noisy labeled data [44]. Notably, even expert annotated data is susceptible to label noise, given the difficulty of the data to annotate. A survey by Song et. al.[33] revealed that the number of corrupt labels in real-world datasets ranges between 8.0% to 38.5%. The authors demonstrate that reducing label noise and creating cleaned data can improve the accuracy of models. To address the issue of noisy labels, an approach known as "repeated-labeling" has been proposed. Repeated-labeling means to obtain annotations from multiple annotators/coders

for the same data entry, such as an image. More specifically: For a set of images $\{x_i\}_{i=1}^{N}$ multiple annotators create noisy labels $\{\tilde{y}_i^r\}_{i=1,\dots,N}^{r=1,\dots,R}$, with $\tilde{y}_i^r$ being the label assigned from annotator $r$ to image $x_i$, but without a ground truth label $\{y_i\}_{i=1,\dots,N}$ [34].

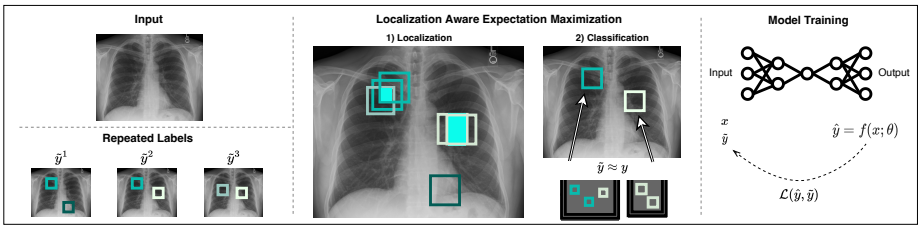


**Fig. 2.** Left: Original input image featuring three separate annotations by distinct annotators. Center: Application of the LAEM aggregation method to the three annotations, yielding an approximate ground truth. Right: Aggregated ground truth utilized during the training process.

Methods for mitigating the negative effect of label noise via repeated labeling can be divided into two categories [18,34]: (a) *two-stage* approaches [39,8] and (b) *one-stage* or *simultaneous* approaches [15,12,11]. Two-stage approaches aim to approximate the ground truth prior to training, a process known as ground truth estimation or ground truth inference [45], as depicted in Figure 2; a straightforward approach is to compute a majority vote. Following label aggregation, the model is trained in a regular fashion. Two-stage approaches offer the benefit of being compatible with commonly used model architectures. On the other hand, simultaneous approaches attempt to integrate repeated labels directly into the training process. In any case, the primary objective of both strategies is to achieve robust and accurate results by leveraging the repeated labeled data to the fullest extent possible. Doing so is crucial to justify the additional annotation efforts. Lastly, to enable the use of established performance metrics, such as those employed in the COCO object detection dataset (mAP) [20], a ground truth estimation step is essential for the validation and test sets. While simultaneous approaches can more effectively utilize repeated labels, they are not intended to execute the necessary aggregation step required to generate the unique ground truth estimate [34]. Consequently, reliable approximation methods are indispensable for evaluation purposes.

Object detection and instance segmentation require both localization and classification, which means that existing methods for repeated labels that are used for classification tasks such as image classification or named entity recognition are not applicable [28]. That is, the available selection of ground truth inference methods is limited. Furthermore, the creation of bounding box or polygonal annotations is expensive [10] and reduces the number of datasets with repeated labels available for evaluating ground truth inference methods [35,23]. However,

we deliberately avoid using synthetic data and focus on real datasets. Our contributions are as follows:

1. We propose a localization algorithm that enables the use of existing ground truth estimation methods such as majority voting or expectation maximization for instance-based recognition tasks and evaluate it extensively with existing methods [32,18].
2. We introduce a comparative analysis of ground truth inference methods that highlights their properties and limits.
3. We conduct ablation studies to analyze the costs associated with creating repeated annotations, and what to do when the amount of available annotated data is limited.
4. We introduce an extension for the TexBiG dataset [35] in the form of a test subset, wherein each of the 200 test images has been annotated by five expert annotators. Utilizing our aggregation method, we establish a unique approximation of the ground truth, which will serve as the unknown reference standard on an evaluation server. This approach allows the TexBiG dataset to be used for evaluation of robust learning methods addressing the challenge of noisy labels.

Once released, the link to the evaluation server will be posted on the GitHub repository where the code is hosted: https://github.com/Madave94/gtiod.

## 2 Related Work

To approximate the ground truth, estimation methods make assumptions about the data and task properties as well as the annotation process. Majority Voting (MJV) [14,29,26] assumes correct labels for the majority of training samples and aggregates the labels accordingly:

$$\tilde{y_i} = \begin{cases} 1 \text{ if } (1/R) \sum_r^R = y_i^r > 0.5 \\ 0 \text{ if } (1/R) \sum_r^R = y_i^r < 0.5 \end{cases} \tag{1}$$

In case of a tie, the label is chosen randomly between the tied ones or selected by a super-annotator. On data with high inter-annotator agreement, majority voting can be a straightforward approach to obtain ground truth estimates reasonable quality.

Numerous methods for inferring ground truth rely on the Expectation-Maximization (EM) approach, first introduced by Dawid and Skene [8]. This approach estimates annotator confidence and integrates it into a weighted voting procedure for determining the true label. By considering annotator performance, these methods address the limitations of majority voting, thereby avoiding potential outliers. One notable advancement in this area is the GLAD [39] method, which not only attempts to identify the most probable class but also assesses image difficulty, additional to the annotator confidence. It should be noted, however, that this approach is limited to binary classification tasks [31].

In addition to classification tasks, pixel-wise classification (semantic segmentation) also has existing ground truth inference methods, such as STAPLE [38], SIMPLE [17], and COLLATE [1]. Recent developments in this field have led to approaches that incorporate data difficulty into the estimation process, as seen in a newly developed simultaneous method [11]. Although there are numerous variations of ground truth estimation methods for classification and segmentation tasks, this discussion will focus on methods applicable to object detection and instance segmentation, rather than diving deeper into this area.

For instance-based recognition tasks like object detection and instance segmentation, there is an additional issue to consider – the localization step. During training, methods consisting of a combination of thresholds and non-maximum suppression are used to solve the localization problem and then focus on classification accuracy. While this may work during training, repeated labeling is likely to have more than just a prediction and ground truth pair to match, since multiple annotators might have created multiple labels. Hence, existing methods are not applicable. An existing approach to aggregate annotations for object detection is called Weighted Boxes Fusion (WBF) [32,18], which was used for the VinDr-CXR dataset [23]. WBF focuses on the weighted aggregation within each class, ignoring inter-class disagreements and also not discarding any annotations even with low agreement. This is beneficial in cases where missing a possible case is far more severe then finding too many, such as a task that requires high recall. Apart from this single existing instance-based recognition approach, we are not aware of any other aggregation methods for object detection or instance segmentation.

## 3 Method

In the following section we introduce a novel adaptation of the EM algorithm, *localization-aware expectation maximization* (LAEM), for instance-based recognition tasks. The same localization algorithm can also be used with majority voting, which therefore functions as a baseline. Additionally, we expand the existing weighted boxes fusion technique to encompass weighted mask fusion, which enables its use in instance segmentation and facilitates benchmarking on a broader range of datasets. As extending weighted boxes fusion is not our core contribution it can be found in Appendix 1.

### 3.1 Localization-Aware Expectation-Maximization

Our novel approach adds a localization stage to existing methods like majority voting and expectation maximization, enabling the use of these established methods for instance-based recognition tasks. Thus, the proposed label aggregation process consists of two stages: (1) *localization-stage* and (2) *classification-stage*. Assuming that $R$ annotators have created noisy instance-based labels $\tilde{y}_{ij}^{r}$ for image $x_i$. Subscript $j = 0, ..., M_r$ refers to the single instances annotated by annotator $r$ for image $x_i$. $M_r$ can be zero if no instances were labeled by $r$.

Each instance contains a class $c \in C$ denoted $\tilde{y}^r_{ijc}$. Furthermore, $\tilde{y}^r_{ijb}$ refers to the respective bounding box and $\tilde{y}^r_{ijs}$ to the optional pixel-wise segmentation mask.

**Algorithm 1** Outline of the localization algorithm used for LAEM

**Require:**
$X = \{x_i\}_{i=1,\dots,N}$ ▷ Set of images
$\tilde{Y} = \{\tilde{Y}_i\}_{i=1,\dots,N}$ ▷ Set of noisy labels per image
$S = \{S_i\}_{i=1,\dots,N}$ ▷ Set of annotators per image
$\theta$ ▷ IoU threshold
**for** $i \in X$ **do** ▷ Loop over images
  $\tilde{Y}_i = \{\tilde{y}^1_{i1}, \tilde{y}^1_{i2}, \dots, \tilde{y}^1_{iM_1}, \tilde{y}^2_{i1}, \tilde{y}^2_{i2}, \dots, \tilde{y}^2_{iM_2}, \dots, \tilde{y}^R_{i1}, \tilde{y}^R_{i2}, \dots, \tilde{y}^R_{iM_R}\}$
  $\tilde{Y}^{\text{LAEM}}_i = \varnothing$
  $Q = \{U_k | U_k \in \mathcal{P}(S_i) \wedge |U_k| \geqslant |U_{k+1}| \wedge \lceil |S_i|/2 \rceil \leqslant |U_k|\}$ ▷ Ordered set of annotator combinations
  **for** $U \in Q$ **do** ▷ Loop over annotator combinations
    $\mathrm{L} = \{\tilde{Y}^{k_1}_{i:} \times \cdots \times \tilde{Y}^{k_n}_{i:} | k_1, \dots, k_n \in U \wedge n = |U|\}$ ▷ Possible combinations of labels
    $F = \{u_k | u_k \in L \wedge \theta \leqslant IoU(u_k) \wedge IoU(u_k) \geqslant IoU(u_{k+1})\}$ ▷ Filtered and ordered L
    **for** $k \in N$ **do** ▷ Loop over label combinations
      $K = \{k_1, k_2, \dots, k_n\}$
      **if** $K \cap \tilde{Y}_i = \varnothing$ **then** ▷ Check for label availability
        $\tilde{Y}_i = \tilde{Y}_i \backslash K$ ▷ Remove labels from available labels
        $\tilde{Y}^{\text{LAEM}}_i = \tilde{Y}^{\text{LAEM}}_i \cup aggregate(K)$ ▷ Add aggregated label to accepted labels
      **end if**
    **end for**
  **end for**
**end for**

Algorithm 1 outlines the LAEM approach. The algorithm requires image set $X$, a set of noisy labels $\tilde{Y}$, a set of annotators $S$, and a threshold $\theta$. Looping over the images of the dataset, the power set $\mathcal{P}(\mathcal{S})$ over the annotators is computed. Subsets containing less than half the number of annotators are removed and a descending order is enforced onto the set. It subsequently iterates through the remaining ordered subsets of annotators and computes the Cartesian product between the respective annotators. Each tuple is then ordered and filtered according to threshold $\theta$ based on the intersection over union in its generalized form:

$$IoU = \frac{\bigcap_{r=1}^{R} \tilde{y}^r_{ijb}}{\bigcup_{r=1}^{R} \tilde{y}^r_{ijb}} \tag{2}$$

The remaining set of tuples ordered by descending IoU forms the set of candidate solutions $F$. In case all labels from a candidate tuple are still available, they are aggregated according to an aggregation function and added to the inferred solutions $\tilde{Y}^{\text{LAEM}}_i$. The aggregation function comprises two steps: (1) all classes contained in the original tuple $\tilde{y}^r_{ijc}$ are appended to a list serving as input for expectation maximization or majority voting and (2) the areas of the different candidates $\tilde{y}^r_{ijb}$ are combined according to the union, intersection, or average area of all boxes involved. The average operation is based on the WBF algorithm [32] with uniform weights. If available, the same procedure (for details cf. Appendix 1) is applied to the segmentation masks $\tilde{y}^r_{ijs}$. This concludes the localization stage. In the subsequent classification-stage existing ground truth

inference methods can be applied such as majority voting or expectation maximization [8].

### 3.2 Algorithmic Design Choices

Our algorithm is designed in a divide-and-conquer manner. Firstly, we prioritize localization, effectively reducing the problem to a classification task for each matched instance after localization. This strategy consequently facilitates the application of established methods for ground truth inference developed in other fields. We always prefer a localization match with more annotators to maximize consensus. If a localization match involving all available annotators cannot be found given the threshold value $\theta$, we ensure successive reduction to potentially prefer the next largest number of annotators. This approach first guarantees localization quality, and only upon establishing matched areas based on their localizations do we aggregate the classes. The algorithm is parameterized by the threshold value $\theta$, which can be adjusted to enforce stricter localization quality and also control the order in which instances are matched. Though this heuristic solution may not provide an optimal outcome for larger problem sizes (e.g., numerous instances on a single image), when an image exhibits high agreement among annotators, a consensus area can be aggregated, and the class of this area can be unambiguously determined.

One advantage of the Expectation-Maximization (EM) approach is that assignment is unambiguous. The confidence calculated during the EM algorithm serves as a tie-breaker, a benefit not present with Majority Voting (MJV). Furthermore, fitting the EM algorithm is efficient; following localization matching, no further areas are calculated, and only the solutions $\tilde{Y}_i^{\mathrm{LAEM}}$ are considered along with their classes.

While localization fusion functions, such as union or intersection, are available and applicable for training data, the intended use for test data within the context of LAEM (Localization-Aware Expectation-Maximization) primarily involves the averaging fusion function. This approach enables a balanced aggregation of areas across different annotators. Additionally, this method is also utilized to aggregate test data as required for the TexBiG dataset [35].

### 3.3 Comparative Analysis

In Table 1, we present a comparative analysis of the four available ground truth inference methods for instance-based recognition tasks, distinguished by their respective characteristics and properties. Each method is described based on eight distinct features relevant to ground truth estimation methods. A noteworthy difference between LAEM and MJV, as compared to WBF and its adaptation (detailed in Appendix 2), is the handling of instances that lack consensus among annotators. Figure 1 and Appendix 3 illustrates the aggregation processes of MJV, LAEM, and WBF on a few specific images, serving as practical examples. This illustrations reveal that MJV and LAEM tend to find consensus instances, resulting in a final image that appears as if a single annotator has labelled the

| Methods | LAEM | MJV | WBF | WBF+EM |
|---|---|---|---|---|
| Assignment | 1) Localization 2) Classification | 1) Localization 2) Classification | Localization only | Localization only |
| Low agreement | Discard annotation | Discard annotation | Keep annotation | Keep annotation |
| Edge cases | Use confidence | Randomized | – | – |
| Localization fusion | Union / average / intersection | Union / average / intersection | Averaging | Weighted averaging |
| Annotator confidence | ✓ | × | × | ✓ |
| Handling missing data | ✓ | ✓ | ✓ | ✓ |
| Dataset characteristic | Balanced | Precision oriented | Recall oriented | Recall oriented |
| Data dependence | × | × | × | × |

**Table 1.** Comparison table for the characteristics and properties of the different ground truth inference methods. MJV and LAEM both use the novel localization algorithm.

image. In contrast, the WBF image is relatively cluttered with overlapping instances. This discrepancy arises because WBF merges areas of the same class that significantly overlap but does not discard any annotation. This is also the case, for instances where two annotators found the same instance area but disagreed on the class, resulting in more instances overall. Although this property might be beneficial for a high-recall scenario − where missing an instance is more detrimental than detecting multiple false positives − it is not ideal for many applications. It's important to note that none of the current methods incorporate data dependence, a feature described in several state-of-the-art ground truth estimation methods for semantic segmentation [17,38,1].

## 4 Experimental Results

In our preliminary experiment, we scrutinized the influence of annotation budget size by exploring scenarios in which repeated labels might be preferred over single labels. This ablation study was designed to determine the optimal use of a restricted annotation budget, i.e., whether it is more beneficial to train with a larger volume of noisy labels or a reduced set of refined labels. This experimental analysis was conducted using two separate datasets.

In our subsequent investigation, we assessed the effect of annotator selection on model performance by deliberately excluding certain annotators from the labeling process. This enabled us to emulate the potential impact of annotator selection on model performance and to probe the influence of proficient and suboptimal annotators on model output.

Our final experiment, which is detailed in Appendix 2, was actually conducted first since it influenced the choice of the aggregation method used for the training data. However, this experiment is not the main focus of this publication.

### 4.1 Set-Up

To the best of our knowledge, there are only two datasets available that contain repeated labels for object detection and instance segmentation, respectively: the VinDr-CXR [23,18] dataset and the TexBiG [35] dataset. We focus solely on these two datasets and do not make use of any synthetic data.

**VinDr-CXR dataset.** This dataset comprises 15,000 training and 3,000 test chest X-ray images, each of which was annotated by three annotators and five annotators, respectively. With a total of 36,096 instances in the training dataset, the dataset can be considered sparsely annotated, with in average 0.8 instances per image[1]. The dataset consists of 14 different classes that were annotated by 17 radiologists [22]. Using the agreement evaluation method presented in [35] describing the data quality, the K-$\alpha$ (Krippendorff's alpha) is 0.79. However, since only 29.3% of the images in the dataset contains any annotations at all, another K-$\alpha$ was calculated for this reduced subset, resulting in a K-$\alpha$ value of 0.29. This indicates that while annotators largely agree in cases where no anomaly is present, there is significant disagreement in cases that contain instances.

**TexBiG dataset.** Recently published [35], the TexBiG provides labels for document layout analysis, similar to the PubLayNet [46] or DocBank [19] datasets. It covers 19 classes for complex document layouts in historical documents during a specific time period, and in the version used here, the training data contains 44,121 instances, the validation data 8,251 instances and the test data 6,678 instances. While the total number of instances is larger as in the VinDr-CXR dataset, there are only 2,457 images in total, 1,922 in the training set, 335 in the validation set and 200 in the test set. Due to the iterative creation process of the dataset, the number of repeated labels is different depending on the sample. An agreement value was used per image to evaluate which samples were to be annotated again. For each image, two annotators were assigned, and in case the agreement value was low after the first iteration, an additional annotator was added to that specific sample. This was done until a maximum of four annotators per sample. In the combined validation and training set, 34 images were annotated by 4 annotators, 336 by at least 3 annotators (including the 34 from before), and 2,257 by at least 2 annotators. We created an additional test set with 5 annotators for 200 newly selected images from the same domain, in accordance with the guideline provided by the authors [35]. We plan to publish this test-set for benchmarking purposes on an evaluation server. The TexBiG dataset is more densely annotated, with 10.7 instances per image, which is 13 time more than the VinDr-CXR dataset. Furthermore, the K-$\alpha$ for the TexBiG training dataset is higher with 0.93.

Comparing the two datasets we find that they represent two opposing marginal cases with one dataset having high-agreement and dense annotations, while the other one has a low-agreement and sparse annotations. However, a more balanced dataset is missing.

[1] Computed by dividing the number of instances by the product of the number of images and the number of annotators.

**Architecture choice.** Regarding the architecture choice, we aimed to find a well-performing and stable choice, rather than aiming for state-of-the-art results since we wanted to focus on comparing the ground truth inference methods and ablation studies on different tasks. For the VinDr-CXR dataset, we tested various architectures including different anchor-based two-stage detectors like Faster R-CNN [27], Cascade R-CNN [2] and Double Head R-CNN [40], and additionally, the transformer-based Detection Transformer (DETR) [3]. After thorough investigation, we found that the Double Head R-CNN performs stably and with reasonable results. Therefore, we selected this architecture for our experiments. On the TexBiG dataset, we tried several instance segmentation models like Mask R-CNN [13], Cascade Mask R-CNN [2] and DetectorRS [24], as well as the Mask2Former [6] as a transformer-based architecture. In this case, DetectoRS yielded the most stable performance, and we continued our experiments with this model. We extended MMDetection [4] with our implementation, and the code is available on GitHub under https://github.com/Madave94/gtiod.

### 4.2 Annotation Budget Ablation

When working on a deep learning project, data is often a limiting factor, and resources must be carefully allocated to create task-specific data. Even when there are ample resources, maximizing the value of those resources is important. In the context of human-annotated data, this concept is referred to as the "annotation budget", which represents the available number of images or instances that can be labeled by a pool of annotators within their available time. The question then becomes, "How can a limited annotation budget be best utilized?" One approach is to prioritize annotating as many different images as possible to cover a broad range of cases within the application domain. However, this approach comes with the risk of introducing more noisy labels due to the inherent variability in annotator performance. Alternatively, creating repeated labels may be more beneficial to improve the quality of the annotations. Ultimately, the decision between prioritizing *quantity versus quality* of labels must be carefully weighed and considered in the context of the project goals and available resources.

In the two ablation studies presented in Table 2 and 3, we compare the performance of different annotation budgets, which refer to the available number of images or instances that can be labeled with the pool of annotators and their available time. The splits used in the studies represent three different cases: (1) Only single annotated labels are available, which are more prone to label noise. (2) A mix of repeated labels and single annotated labels is available. Multiple splits may have this property. (3) Maximum label repetition, where no or very few single annotated labels are available, resulting in significantly less training data. To reduce randomization effects, we create five different random versions for each split and compute their mean and maximum results.

Our results show that the TexBiG dataset quickly reached a data saturation point, suggesting potential benefits from employing multi-annotator learning methods to better utilize repeated labels. Conversely, the VinDr-CXR dataset

| Split | Budget rel. | Budget abs. | Averaged AP | Averaged $AP^{bb}$ | Maximum AP | Maximum $AP^{bb}$ |
|---|---|---|---|---|---|---|
| 1922×2 | 100% | 3844 | 41.9 | 47.5 | 43.3 | 48.7 |
| 966×2<br>966×1 | 75% | 2883 | 42.4 | 47.9 | 42.8 | 48.4 |
| 1922×1 | 50% | 1922 | **42.7** | **48.3** | 43.4 | **48.8** |
| 641×2<br>640×1 | 50% | 1922 | 42.4 | 47.7 | **43.9** | **48.8** |
| 966×2 | 50% | 1922 | 41.1 | 46.2 | 42.8 | 47.8 |
| 30×4<br>243×3<br>1073×1 | 50% | 1922 | 41.9 | 47.3 | 43.0 | 48.2 |
| 30×4<br>243×3<br>536×2<br>1×1 | 50% | 1922 | 39.7 | 45.0 | 41.0 | 46.5 |

**Table 2.** Ablation study on the TexBiG dataset using a limited annotation budget. The results are in $mAP$@[.5 : .95], show that multi annotator learning methods are required to justify repeated labels. However, even without multi annotator methods the performance loss using repeated annotations is marginal.

| Split | Budget rel. | Budget abs. | Avg. AP | Max. AP |
|---|---|---|---|---|
| 15,000×2 | 66.6% | 30,000 | 14.8 | 15.0 |
| 10,000×3 | 66.6% | 30,000 | 14.7 | 14.9 |
| 15,000×1 | 33.3% | 15,000 | 13.4 | 13.9 |
| 10,000×1<br>2,500×2 | 33.3% | 15,000 | **13.6** | 14.1 |
| 7,500×2 | 33.3% | 15,000 | 13.5 | 13.8 |
| 3,000×3<br>3,000×2 | 33.3% | 15,000 | **13.6** | **14.3** |
| 5,000×3 | 33.3% | 15,000 | 13.4 | 14.0 |

**Table 3.** Ablation study on the VinDr-CXR dataset using a limited annotation budget. The results are in $mAP_{40}$ as provided by the leaderboard.

showed improved performance with higher budgets, indicating that more data helps performance in scenarios with noisy, low-agreement labels.

Both datasets demonstrate that moderate inclusion of repeated labels does not adversely impact performance, with mixed splits achieving peak results at their lowest budgets. These findings highlight the value of repeated annotations, which not only increase label reliability, but also allow for efficient use of multi-annotator learning methods. Remarkably, the opportunity costs for creating such repeated labels seem negligible.

Our findings suggest that higher fragmentation in annotator splits could lead to reduced performance, possibly due to enhanced intracoder consistency. Moreover, the influence of split distribution appears prominent only when the annotation budget is limited. Identifying a systematic relationship between split distribution and performance, thereby suggesting optimal splits before the annotation process, could be a promising future research direction.

The overall takeaway is that multiple annotations may not always yield significant advantages, yet in scenarios with a constrained annotation budget, they could prove beneficial. Determining which cases fall into each category remains an open challenge.

### 4.3 Leave-One-Out Annotator Selection

Table 4 displays the results of a final experiment conducted on the TexBiG dataset. To create four groups of annotators, each group consisting of one to three individuals, annotations were distributed unevenly among them, resulting in groups of different sizes. Subsequently, each group was left out of the training process, while the remaining three groups were used to train the model. This approach led to a smaller training set. Surprisingly, the experiment showed that when the largest group, denoted as **B**, was excluded, leaving only 61.6% of the annotations available, the model's performance reached its peak. This outcome underscores the importance of selecting precise annotators in the training process, since less precise ones may introduce noisy labels that can hinder performance. However, it is challenging to identify precise annotators before the annotation process, as there is no data available to determine their level of precision.

| Left out group | Left out images | | Left out annotations | | Performance | |
|---|---|---|---|---|---|---|
| | rel. | abs. | rel. | abs. | AP | $AP^{bb}$ |
| Group A | 25.1% | 1,040 | 26.8% | 11,810 | 42.4 | 47.1 |
| Group B | 29.5% | 1,225 | 38.4% | 16,932 | **44.1** | **49.8** |
| Group C | 25.7% | 1,067 | 18.2% | 8,017 | 42.6 | 48.1 |
| Group D | 19.7% | 815 | 16.7% | 7,362 | 43.1 | 48.3 |

**Table 4.** Choosing the right annotator? If annotators are not in the group what would happen to the results? Splits are unequal, due to the annotation distribution.

## 5 Conclusion

Our results indicate the potential benefits of repeated labels which seem to be contingent on several factors. The identified key factors are the balance between (1) the complexity or variation in the dataset and its corresponding task difficulty, (2) the variability in annotation depending on inter-annotator consistency and annotator proficiency, and (3) the constraints of the annotation budget. This interaction suggests the existence of an 'optimal range' for image annotation strategy. For instance, datasets with high variance and low annotator consistency may benefit from multiple annotations per image, while in cases with low image variation and high annotator consistency, many images annotated once might suffice. This balancing act between data and annotation variation could guide decisions when choosing between single or multiple annotators per image, given a fixed annotation budget.

However, the utility of repeated labels is substantially hampered due to the lack of multi-annotator-learning approaches for object detection and instance

segmentation. Thus, future work should concentrate on developing methods that bridge this gap between these areas and other computer vision domains like image classification or semantic segmentation.

Lastly, a significant challenge remains regarding the availability of suitable datasets. With limited datasets in the domain and disparities among them, our findings' generalizability remains constrained to the two domains covered in this study. A larger dataset with repeated labels and balanced agreement would be valuable for future research. Synthetic data could be beneficial but pose the risk that models trained on these data may only learn the distribution used to randomly create repeated labels from the original annotations. Thus, creating a suitable dataset remains a formidable task.

**Acknowledgments.** This work was supported by the Thuringian Ministry for Economy, Science and Digital Society / Thüringer Aufbaubank (TMWWDG / TAB).

## References

1. Asman, A.J., Landman, B.A.: Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (collate). IEEE transactions on medical imaging **30**(10), 1779–1794 (2011)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence p. 1–1 (2019). https://doi.org/10.1109/tpami.2019.2956516, http://dx.doi.org/10.1109/tpami.2019.2956516
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 213–229. Springer (2020)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)
7. Cheng, Y., Zhu, J., Jiang, M., Fu, J., Pang, C., Wang, P., Sankaran, K., Onabola, O., Liu, Y., Liu, D., et al.: Flow: A dataset and benchmark for floating waste detection in inland waters. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10953–10962 (2021)
8. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics) **28**(1), 20–28 (1979)

9. Feng, D., Wang, Z., Zhou, Y., Rosenbaum, L., Timm, F., Dietmayer, K., Tomizuka, M., Zhan, W.: Labels are not perfect: Inferring spatial uncertainty in object detection. IEEE Transactions on Intelligent Transportation Systems (2021)
10. Gao, J., Wang, J., Dai, S., Li, L.J., Nevatia, R.: Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9508–9517 (2019)
11. Gao, Z., Sun, F.K., Yang, M., Ren, S., Xiong, Z., Engeler, M., Burazer, A., Wildling, L., Daniel, L., Boning, D.S.: Learning from multiple annotator noisy labels via sample-wise label fusion. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. pp. 407–422. Springer (2022)
12. Guan, M., Gulshan, V., Dai, A., Hinton, G.: Who said what: Modeling individual labelers improves classification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
14. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical image analysis **65**, 101759 (2020)
15. Khetan, A., Lipton, Z.C., Anandkumar, A.: Learning from noisy singly-labeled data. arXiv preprint arXiv:1712.04577 (2017)
16. Khodabandeh, M., Vahdat, A., Ranjbar, M., Macready, W.G.: A robust learning approach to domain adaptive object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 480–490 (2019)
17. Langerak, T.R., van der Heide, U.A., Kotte, A.N., Viergever, M.A., Van Vulpen, M., Pluim, J.P.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). IEEE transactions on medical imaging **29**(12), 2000–2008 (2010)
18. Le, K.H., Tran, T.V., Pham, H.H., Nguyen, H.T., Le, T.T., Nguyen, H.Q.: Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. arXiv preprint arXiv:2203.10611 (2022)
19. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: Docbank: A benchmark dataset for document layout analysis. arXiv preprint arXiv:2006.01038 (2020)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
21. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484 (2019)
22. Nguyen, D.B., Nguyen, H.Q., Elliott, J., KeepLearning, Nguyen, N.T., Culliton, P.: Vinbigdata chest x-ray abnormalities detection (2020), https://kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection
23. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)
24. Qiao, S., Chen, L.C., Yuille, A.: Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10213–10224 (2021)
25. Ramamonjison, R., Banitalebi-Dehkordi, A., Kang, X., Bai, X., Zhang, Y.: SimROD: A Simple Adaptation Method for Robust Object Detec-

tion. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3550–3559. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00355, https://ieeexplore.ieee.org/document/9711168/

26. Raykar, V.C., Yu, S., Zhao, L.H., Jerebko, A., Florin, C., Valadez, G.H., Bogoni, L., Moy, L.: Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proceedings of the 26th Annual international conference on machine learning. pp. 889–896 (2009)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
28. Rodrigues, F., Pereira, F.: Deep learning from crowds. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
29. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 614–622 (2008)
30. Sheng, V.S., Zhang, J.: Machine learning with crowdsourcing: A brief summary of the past research and future directions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9837–9843 (2019)
31. Sinha, V.B., Rao, S., Balasubramanian, V.N.: Fast dawid-skene: A fast vote aggregation scheme for sentiment classification. arXiv preprint arXiv:1803.02781 (2018)
32. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021)
33. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022)
34. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N.: Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11244–11253 (2019)
35. Tschirschwitz, D., Klemstein, F., Stein, B., Rodehorst, V.: A dataset for analysing complex document layouts in the digital humanities and its evaluation with krippendorff's alpha. In: DAGM German Conference on Pattern Recognition. pp. 354–374. Springer (2022)
36. Wang, X., Huang, T.E., Liu, B., Yu, F., Wang, X., Gonzalez, J.E., Darrell, T.: Robust object detection via instance-level temporal cycle confusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9143–9152 (2021)
37. Wang, Z., Li, Y., Guo, Y., Fang, L., Wang, S.: Data-uncertainty guided multi-phase learning for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4568–4577 (2021)
38. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)
39. Whitehill, J., Wu, T.f., Bergsma, J., Movellan, J., Ruvolo, P.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in neural information processing systems **22** (2009)

40. Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., Fu, Y.: Rethinking classification and localization for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10186–10195 (2020)
41. Wu, Z., Suresh, K., Narayanan, P., Xu, H., Kwon, H., Wang, Z.: Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1201–1210 (2019)
42. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1369–1378 (2017)
43. Zhang, H., Wang, J.: Towards Adversarially Robust Object Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 421–430. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00051, https://ieeexplore.ieee.org/document/9009990/
44. Zhang, Z., Zhang, H., Arik, S.O., Lee, H., Pfister, T.: Distilling effective supervision from severe label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9294–9303 (2020)
45. Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R.: Truth inference in crowdsourcing: Is the problem solved? Proceedings of the VLDB Endowment **10**(5), 541–552 (2017)
46. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1015–1022. IEEE (2019)
47. Zhu, X., Wu, X.: Class noise vs. attribute noise: A quantitative study. The Artificial Intelligence Review **22**(3), 177 (2004)

## Appendix 1

This section presents our adaptation of the weighted box fusion (WBF) technique, tailored specifically for instance segmentation as a weighted mask fusion (WMF).

In their study, [18] propose a method for combining annotations from multiple annotators using a weighted box fusion [32] approach. In this method, bounding boxes are matched greedily only with boxes of the same class, and no annotations are discarded. The WBF algorithm fuses boxes that exceed a specified overlap threshold, resulting in new boxes that represent the weighted average of the original boxes. The approach also allows for inclusion of box confidence scores and prior weights for each annotator.

To extend the WBF method for instance segmentation, we introduce an option to fuse segmentation masks, which involves four steps: (1) calculating the weighted area and weighted center points from the different masks, (2) compute the average center point and average area from the selected masks, (3) determining the closest center point of the original masks to the weighted center point and selecting this mask, and (4) dilating or eroding the chosen mask until the area is close to the averaged area. The resulting mask is used as the aggregated segmentation mask and is also used as the averaging operation during the aggregation for LAEM and MJV with uniform weight.

Moreover, we integrate the WBF approach with LAEM, yielding WBF+EM. This integration involves assessing annotator confidence using LAEM, and subsequently incorporating it into the WBF method to produce weighted average areas instead of simply averaged areas. While the differences between LAEM and WBF might seem subtle, WBF+EM offers a more thorough approach to annotator fusion. This modification is relatively minor, and its impact is modest, as corroborated by our experiments delineated in Appendix 2.

## Appendix 2

In this experiment, we carried out a comparative analysis of different ground truth inference methods. To do this, we separated the annotations for training and testing, and created various combinations of train-test datasets using the available ground truth estimation methods. Afterward, a model was trained on these combinations. The results from this experiment reveal how aggregation methods can impact the performance of the trained models and show how these outcomes can vary based on the specific combination of training and testing aggregation used.

Tables 5 and 6 present the application of various ground truth estimation methods on repeated labels. In the TexBiG dataset, each method is employed to aggregate the labels of both training and test data, and all possible train-test combinations are learned and tested to perform a cross comparison of the different ground truth inference methods, as shown in Table 5. The hyperparameter for the area combination is denoted as $\cup$ for union, $\mu$ for averaging and $\cap$ for intersection. Additionally, the plain repeated labels, without any aggregation, are

| DetectoRS TexBiG | | | | Training | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | RL | MJV | | | LAEM | | | WBF | |
| | | | | | $\cup$ | $\mu$ | $\cap$ | $\cup$ | $\mu$ | $\cap$ | base | EM |
| Test | MJV | $\cup$ | AP | 32.5 | 34.5 | 30.4 | 25.5 | **35.1** | 29.1 | 28.4 | 30.2 | 30.7 |
| | | | $AP^{bb}$ | 34.7 | 36.6 | 34.0 | 29.9 | **37.5** | 32.7 | 33.4 | 34.4 | 33.5 |
| | | $\mu$ | AP | 41.9 | **43.9** | 39.9 | 35.8 | 43.6 | 40.8 | 35.0 | 41.5 | 41.9 |
| | | | $AP^{bb}$ | 45.6 | **48.2** | 44.6 | 41.4 | 47.9 | 44.5 | 40.2 | 45.5 | 46.2 |
| | | $\cap$ | AP | 44.2 | 41.7 | **46.6** | 45.0 | 43.2 | 45.3 | 45.2 | 46.0 | 44.9 |
| | | | $AP^{bb}$ | 49.6 | 47.9 | **51.4** | 49.9 | 49.3 | 50.6 | 49.5 | 51.0 | 49.2 |
| | LAEM | $\cup$ | AP | 31.5 | 34.9 | **44.4** | 25.9 | 33.6 | 30.5 | 26.8 | 29.8 | 31.1 |
| | | | $AP^{bb}$ | 33.6 | 36.5 | **48.9** | 30.5 | 35.8 | 33.8 | 31.4 | 33.0 | 34.6 |
| | | $\mu$ | AP | 41.1 | 42.5 | 40.5 | 34.9 | **43.6** | 40.2 | 35.9 | 40.7 | 40.4 |
| | | | $AP^{bb}$ | 44.8 | 46.8 | 44.6 | 41.9 | **47.7** | 44.3 | 41.7 | 45.5 | 45.1 |
| | | $\cap$ | AP | 43.5 | 40.8 | 43.0 | 45.0 | 41.6 | 44.1 | 44.0 | 45.0 | **45.1** |
| | | | $AP^{bb}$ | 49.8 | 46.0 | 48.5 | 49.5 | 46.9 | 48.9 | 48.0 | 49.5 | **50.3** |
| | WBF | base | AP | 36.1 | **38.0** | 34.8 | 33.4 | 37.3 | 30.3 | 32.7 | 34.9 | 36.9 |
| | | | $AP^{bb}$ | 38.8 | **41.6** | 38.0 | 37.4 | 40.2 | 33.6 | 37.8 | 38.5 | 40.3 |
| | | EM | AP | 38.1 | **39.9** | 34.9 | 32.4 | 39.8 | 36.5 | 32.8 | 36.3 | 35.9 |
| | | | $AP^{bb}$ | 40.6 | **42.8** | 38.3 | 36.4 | 42.8 | 40.2 | 37.6 | 40.0 | 39.3 |
| | Mean | | AP | 38.6 | 39.5 | 39.3 | 34.7 | **39.7** | 37.1 | 35.1 | 38.1 | 38.4 |
| | | | $AP^{bb}$ | 42.2 | 43.3 | **43.5** | 39.6 | **43.5** | 41.1 | 40.0 | 42.2 | 42.3 |

**Table 5.** Cross-Validation of ground truth inference combinations between training and test data, for the DetectoRS with a ResNet-50 backbone on the TexBiG dataset. Showing the $mAP@[.5 : .95]$ for instance masks and bounding boxes. Union is represented by $\cup$, intersection by $\cap$ and averaging by $\mu$. RL denotes training conducted on un-aggregated noisy labels. The two rows on the bottom show how the training methods perform on average.

compared with the different aggregated test data. Our findings reveal that on a high-agreement dataset, weighted boxes fusion does not perform well. This could be attributed to the inclusion of most annotations by WBF, whereas in cases with high agreement, it is more desirable to exclude non-conforming instances. Majority voting and localization-aware expectation maximization perform similarly; however, LAEM provides a more elegant solution for addressing edge cases. Calculating the annotator confidence, as performed in LAEM, is highly advantageous. However, in rare cases, spammer annotators could potentially circumvent annotation confidence by annotating large portions of simple examples correctly but failing at hard cases. Such cases would result in a high confidence level for the spammer, potentially outvoting the correct annotators on challenging and crucial cases.

The main performance differences between MJV and LAEM arise due to the application of the three different combination operations – union, averaging, and intersection. Combining areas by taking their union results in larger areas, making it easier for a classifier to identify the respective regions. Analysis of the mean results of the training methods reveals that both MJV+$\cup$ and LAEM+$\cup$

| Double Head R-CNN VinDr-CXR | Training | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | RL | MJV | | | LAEM | | | WBF | |
| | | $\cup$ | $\mu$ | $\cap$ | $\cup$ | $\mu$ | $\cap$ | base | EM |
| private LB | **16.2** | 15.2 | 14.6 | 14.3 | 14.9 | 15.0 | 14.7 | 13.7 | 14.3 |

**Table 6.** Comparing results with the private Kaggle leaderboard [22] for the VinDr-CXR dataset using the double headed R-CNN at $mAP_{40}$. Union is represented by $\cup$, intersection by $\cap$ and averaging by $\mu$. RL denotes training conducted on un-aggregated noisy labels.

exhibit the highest performance across various test configurations. On the contrary, methods parameterized with intersection $\cap$ yield the lowest mean results. Training with repeated labels without any aggregation yields results similar to training with aggregated labels. However, while it may be generally feasible to train with noisy labels, the performance is slightly dampened. Since the test data aggregation method is LAEM-$\mu$ as described in Section 3.2, the best performing training method LAEM-$\cup$ is choosen as the aggregation method for the training data in the experiments shown in Section 4.2 and 4.3.

For the VinDr-CXR dataset, a smaller, similar experiment is performed as shown in Table 6. As the Kaggle leaderboard already provides an aggregated ground truth and labels are unavailable, only the training data are aggregated. Our findings indicate that training with plain repeated labels leads to higher results. Given the low agreement of the dataset, training with repeated labels may be seen as a form of "label augmentation." Interestingly, the methods used to aggregate the test data, such as WBF, do not outperform the other methods. However, ground truth estimation methods are not designed to boost performance but rather to provide a suitable estimation for the targeted outcome. Based on these results, the following experiments on VinDr-CXR will be run with the repeated labels for training.

## Appendix 3

This section shows three more comparisons between different ground truth aggregation methods, exemplary on the VinDr-CXR dataset [23]. All of them follow the same structure. Left: the original image with the repeated labels indicated by the different line types. Right: the four smaller images from top left to bottom right are, MJV+$\cap$, LAEM+$\mu$, LAEM+$\cup$ and WBF.

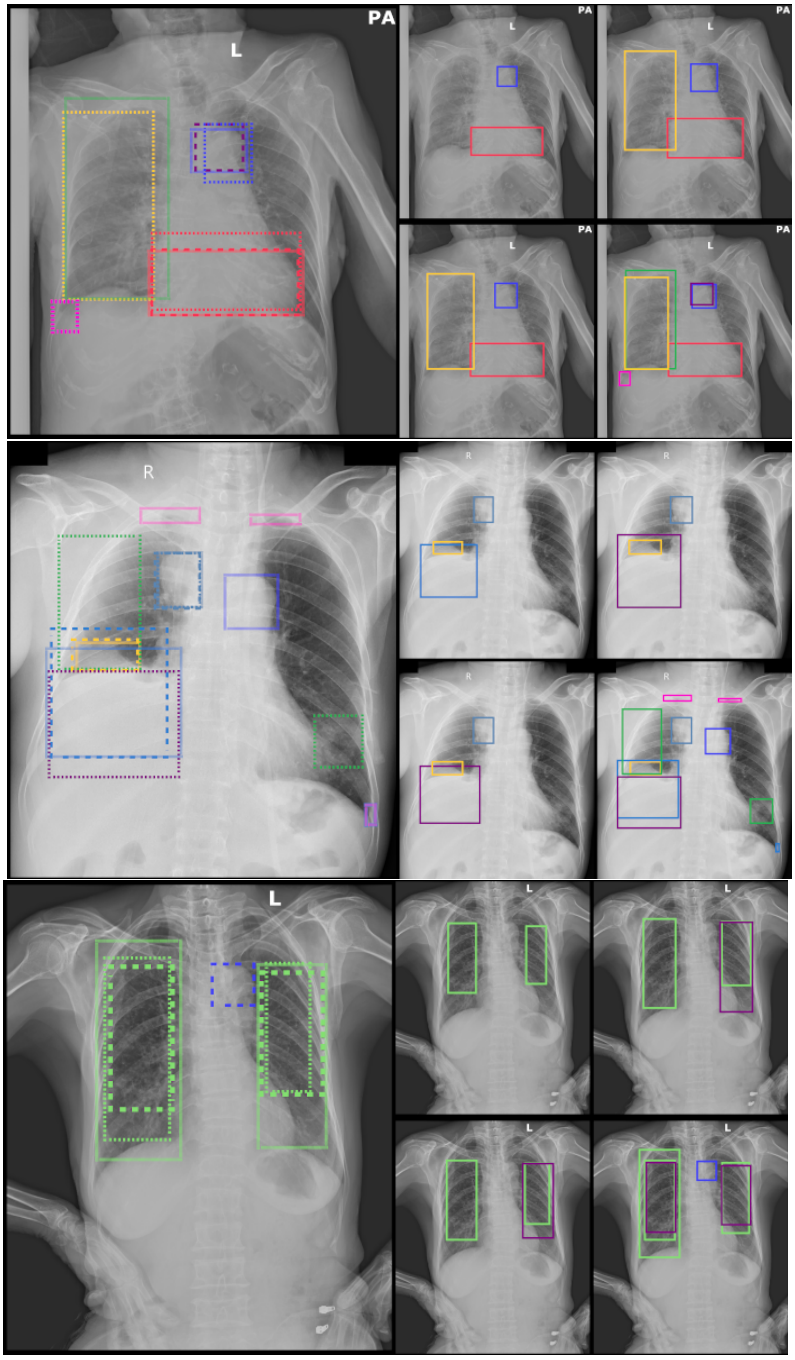


**Fig. 3.** Qualitative results on three test images from the VinDr-CXR. Left: the original image with the repeated labels indicated by the different line types. Right: the four smaller images from top left to bottom right are, MJV+$\cap$, LAEM+$\mu$, LAEM+$\cup$ and WBF.