

Label Convergence: Defining an Upper Performance Bound in Object Recognition through Contradictory Annotations

David Tschirschwitz Volker Rodehorst

Bauhaus-Universität Weimar, Germany

david.tschirschwitz@uni-weimar.de

Abstract

Annotation errors are a challenge not only during training of machine learning models, but also during their evaluation. Label variations and inaccuracies in datasets often manifest as contradictory examples that deviate from established labeling conventions. Such inconsistencies, when significant, prevent models from achieving optimal performance on metrics such as mean Average Precision (mAP). We introduce the notion of “label convergence” to describe the highest achievable performance under the constraint of contradictory test annotations, essentially defining an upper bound on model accuracy.

Recognizing that noise is an inherent characteristic of all data, our study analyzes five real-world datasets, including the LVIS dataset, to investigate the phenomenon of label convergence. We approximate that label convergence is between 62.63-67.52 mAP@[0.5:0.95:0.05] for LVIS with 95% confidence, attributing these bounds to the presence of real annotation errors. With current state-of-the-art (SOTA) models at the upper end of the label convergence interval for the well-studied LVIS dataset, we conclude that model capacity is sufficient to solve current object detection problems. Therefore, future efforts should focus on three key aspects: (1) updating the problem specification and adjusting evaluation practices to account for unavoidable label noise, (2) creating cleaner data, especially test data, and (3) including multi-annotated data to investigate annotation variation and make these issues visible from the outset.

1. Introduction

Machine learning systems can be categorized into three broad phases: (1) dataset creation, (2) model development, and (3) evaluation. Within the computer vision community, there is a significant focus on the second phase—the development of innovative methods to address new challenges or improve results for existing problems. In such

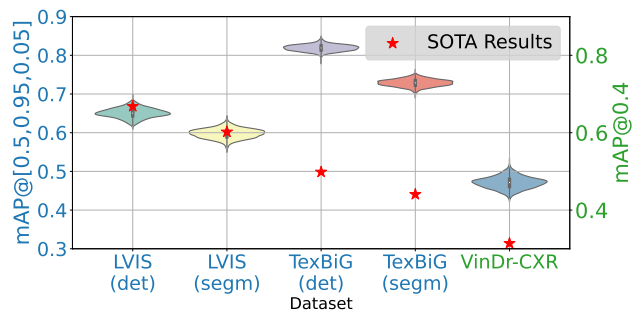


Figure 1. Illustration of the convergence threshold intervals and the respective state-of-the-art (SOTA) results, highlighting how close the data points are to the upper performance bound – at least for LVIS. For the LVIS dataset, the convergence threshold and SOTA results (using pre-trained Co-DETR [42]) are evaluated on the consistency subset of LVIS. The convergence threshold interval for LVIS is created directly using modified mAP@[0.5,0.95,0.05] as described in Section 4.1. For the TexBiG and VinDr-CXR datasets, the convergence threshold interval is inferred using the formula from $K-\alpha$ to mAP introduced in Section 4.2. The TexBiG and VinDr-CXR datasets utilize their respective leaderboard results, with TexBiG using modified mAP@[0.5,0.95,0.05] and VinDr-CXR using the mAP@0.4 metric as used in their respective leaderboards.

cases, the training data, test data, and evaluation criteria remain constant, and the sole focus is on model modifications and adaptations to enhance system performance. In contrast to this model-centric paradigm, there exists an alternative approach known as data-centric AI. This concept was introduced by Andrew Ng [9] and primarily concentrates on the first phase. This strategy involves maintaining a constant model while systematically enhancing the quality of the data. Within this data-centric approach, improvements focus on aspects such as label consistency, the guidelines

Name	Images	Classes	Instance	# Rat./Img.	Masks	Guideline
LVIS Cons. Subset [11]	5,000	1203	100,480	2.00	Yes	Same
COCO Reann. [21]	80,067	5	1,022,716	2.00	Yes	Different
Open Images Reann. [21]	4,773	5	31,198	2.00	No	Different
TexBiG [39]	2,257	19	53,623	2.16	Yes	Same
VinDr-CXR [27]	15,000	14	36,557	3.00	No	Same

Table 1. Datasets with repeated labels, that enable the analysis of annotation variation. TexBiG and VinDr-CXR have repeated labels for the full dataset. For LVIS, only 5,000 and/or the 19,809 validation data are used because they were annotated by multiple persons for consistency measurement. COCO and Open Images use the original and reannotated versions combined to identify label variation, but use different guidelines and have some inconsistencies in the organization of the annotations as described in Appendix A.

or conventions followed by annotators¹, enhancing domain coverage, adjusting the dataset size, and identifying errors, among others. Both methods aim to boost performance, either through advancements in data quality or through model optimization.

In this study, we adopt an alternate perspective by including an additional “zeroth” phase – problem specification (0). Our focus is on both the problem specification (0) and the evaluation (3), and we frame our investigation around the following central question:

“How can we estimate the intrinsic performance threshold for models and data given annotation variation in human labels such as noise or uncertainty?”

This study introduces the concept of *label convergence*, which hypothesizes that the performance limitations of a dataset arise from internal label inconsistencies. These inconsistencies may stem from variations in labeling conventions, annotator variability, or outright errors. Many of these challenges originate in the problem specification phase, where datasets typically assume the provided labels are “gold standard”, representing a singular ground truth. However, for complex data with unavoidable label noise, perfect ground truth is often unattainable. In such cases, label convergence serves as a measure of the inherent ambiguity in both problem specification and evaluation. Before delving into this topic, we first review the significance of labels in machine learning.

The issues with annotations are referred to by several terms: noisy labels, human label variation, annotation errors, or uncertain labels. We will refer to them as (annotation/label) *variations*. Despite the different terminologies, they all denote the same underlying concept where the mapping from a feature vector to a label is not uniquely defined. This phenomenon can be expressed for a set of images, denoted by $\{x_i\}_{i=1}^N$, where an annotator has produced a set of

¹We consider the guideline as the instructions given to the annotators and the annotation convention, the interpretation of annotators made, which in the best case exactly matches the guideline if the guideline is unambiguous.

labels $\{\tilde{y}_{ij}\}_{j=0}^M$ for each image i . Here, j refers to individual instances within an image, such as a bounding box or a segmentation mask, that are intended to represent specific objects or segments. These annotated labels aim to approximate the true but unknown labels y_{ij} [37, 38].

The detrimental effects of label variation on the training process have been extensively studied, with results indicating significant performance degradation [4], negative effects on the training process [22], and limitations on the accuracy of learning algorithms [37]. It has been highlighted that among various types of noise, label variation is particularly damaging. This assertion is supported by evidence of a substantial performance difference between training on clean versus varying data, as demonstrated in the research by Song *et al.* [35]. Additionally, a study by Xu *et al.* [40] revealed that missing 20% of the instances for training can lead to a drop in performance of about 5 percentage points (from 70 to 65 mAP on PASCAL VOC 2007 [10]). This form of noise is of particular concern because deep neural networks have a significant memory capacity, which can lead to overfitting to varying labels if they are prevalent in the training dataset [41]. To reduce the impact of varying labels, it may be preferable to use smaller capacity models but train on cleaner labels [28, 35], especially since performance improvements are only logarithmic with increasing dataset size [36].

According to research referenced by Song *et al.* [35], the prevalence of noisy or incorrect labels in “real-world datasets” is estimated to range from 8.0% to 38.0%. These results primarily concern the domain of image classification. However, to the best of our knowledge, and as supported by Schubert *et al.* [34] and Agnew *et al.* [1], the extent of label variation within object detection datasets remains undetermined.

Training with varying labels has been categorized into four distinct strategies by Plank [32], who specifically addresses annotation variation as human label variation, with a notable emphasis on natural language processing (NLP). These strategies are divided into two main objectives: (1) addressing human label variation through either (1.1) aggre-

gating labels or (1.2) filtering out noise, and (2) leveraging human label variation by either (2.1) learning directly from unaggregated labels or (2.2) integrating gold standard labels with human label variation.

The issue of learning with varying annotations has received attention in the research community, but often adopts a model-centric AI perspective, especially with respect to the test data. This approach typically assumes that the test data are either of sufficient quality, or that any flaws within the test data are deliberately overlooked in order to prioritize model improvements. In this realm, strategies for managing annotation variation often involve the synthetic introduction of variation [4, 18, 20], aiming for mitigation through techniques such as co-teaching [6], soft-label learning [14, 17, 29], or noise filtering [22, 23].

A subset of methods also tackles the real-world label variation that occurs in datasets with “repeated labels” – that is, labels provided by multiple annotators for the same image. Extending the earlier mathematical formulation, such repeated annotations would be denoted by \tilde{y}_{ij}^r , where $r = 1, \dots, R$ represents the index of the annotator. In the context of object detection and instance segmentation, two significant datasets have been introduced: the VinDr-CXR dataset [27] and the TexBiG dataset [39]. Research efforts focused that have shown results on these datasets with real label variations [12, 30, 31, 38] primarily explore label aggregation techniques to approximate the true label.

While the focus of many studies on varying labels has been on enhancing the training process, it is the test data that ultimately determine the system’s utility in practice. For a system to be viable, it’s imperative that the test data comprehensively represent the problem domain, thereby minimizing the discrepancy between controlled (in-vitro) and real-world (in-vivo) conditions [32]. This discussion circles back to the concept of *label convergence* introduced earlier, emphasizing the critical nature of precise and consistent labeling and if that is not further possible a problem specification that accounts for the remaining uncertainties.

When considering the application of neural networks, we distinguish between three types of upper bounds that can affect performance:

1. **Model Convergence:** This refers to the ability of a model to learn from data and generalize to unseen data. Key issues include model bias and variance, which are affected by factors such as optimization techniques, network architecture, and other parameters.
2. **Data Convergence:** This involves the completeness of the dataset. Models typically struggle to learn from data outside their training domain, which affects their ability to generalize. This category also includes considerations of data augmentation techniques.

3. **Label Convergence:** This hypothesis suggests that dataset performance is inherently limited by conflicting labels due to annotation variation. These inconsistencies can significantly hinder the learning process.

In addition to these types of convergence, system performance may also be constrained by hardware limitations or other external factors. Our contributions to the understanding of label convergence in object recognition include:

- We present a method for evaluating the convergence threshold interval using a modified mAP, allowing direct comparison with model results and determining the upper-performance bound. This method is applied to estimate the convergence threshold for the LVIS dataset.
- We extend this method to multiple annotators by correlating the modified mAP with an inter-annotator agreement metric and estimate the convergence threshold for the VinDr-CXR and TexBiG datasets.
- We provide the first analysis of real annotation variation within object detection and instance segmentation, and offer a tangible distribution of this annotation variation across five datasets.

2. Related Work

Previous studies have approached the concept of label convergence indirectly, though not explicitly named as such. These efforts, aimed at understanding the upper bounds of model performance, have generally relied on the analysis of learned models. However, assessing label convergence ideally requires a model-independent approach that focuses solely on label quality.

Borji *et al.* [3] investigated an “Empirical Upper Bound” for object detection models, potentially touching on aspects of the three types of convergence we’ve discussed. They hypothesized an upper bound in mAP determined by the performance of an ideal detector using ground truth bounding boxes and the optimal object classifier. While intriguing, their methodology overlooks localization and relies on learned models, which deviates from a pure label convergence evaluation, which should be independent of any model.

Agnew *et al.* [1] attempted to quantify annotation quality empirically by introducing synthetic variations into the annotations. Their methods include uniform noise for expanding bounding boxes and Gaussian radial noise for polygons. While the objective is to identify a performance ceiling due to contradictory examples, the reliance on synthetic noise and learned models limits the study. These noise patterns, being predictable, are significantly easier to learn by models [24], and the assessment is not fully detached from the original, potentially flawed, annotations.

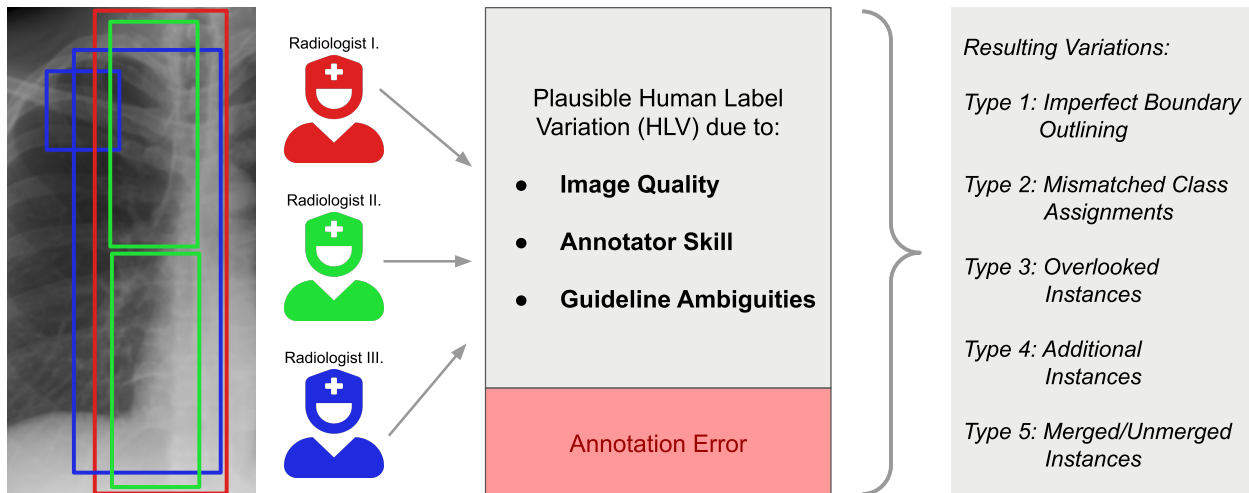


Figure 2. Human label variation causes and annotation errors causing five different types of variations. Our graphic is inspired by Plank [32]. The fifth error type is illustrated between the green and red radiologist.

Ma *et al.* [21] adopted a data-centric AI approach by re-annotating datasets such as MS COCO and Google Open Images. By comparing model performance across various combinations of old and new annotations, they explored the impact of different labeling conventions. While their work points to the interplay between model, data, and label convergence, it again depends on trained models, making it a partial attempt at empirical label convergence evaluation.

In summary, while these studies illuminate challenges of annotation, they fall short of defining label convergence because they rely on learned models and their primary objectives do not align with a pure assessment of label quality.

A promising approach for assessing label convergence is to examine annotation consistency, similar to how performance metrics evaluate model quality. While measuring annotation consistency is common in natural language processing data collections [32], major datasets such as MS COCO [19] lack such consistency metrics. Conversely, datasets like LVIS [11] attempt to measure consistency but use techniques designed to evaluate model predictions against ground truth, such as an F1-score. Even the most common metrics like mAP used for COCO would not work out of the box for several reasons: 1) the ranking of detections in model evaluation relies on confidence scores, which are absent in repeated labeling scenarios because no prediction score is provided, and 2) precision and recall measures are impractical without a definitive count of total detections. Typically, consistency is assessed using statistical measures of inter-annotator agreement, such as Cohen’s Kappa [8] or Krippendorff’s Alpha [13, 15]. However, adapting these methods for object detection or instance segmentation is challenging due to the need to incorporate localization.

Existing methods for semantic segmentation, such as

Ribeiro *et al.* [33], apply Cohen’s Kappa on a per-pixel basis, but overlook the distinction between instances. Similarly, Nassar *et al.* [26] conduct a pixel-wise inter-annotator evaluation using Krippendorff’s Alpha for object detection, again neglecting instance-specific considerations. To our knowledge, the only peer-reviewed approach to evaluating annotation consistency for object detection that considers the instance-based nature has been developed for the TexBiG [39] dataset. It uses a unique method that combines Intersection over Union (IoU) thresholding with graph-based matching to address localization before applying a conventional Krippendorff’s Alpha for assessment.

In addition to assessing the consistency of annotations, several studies have analyzed the types of variations prevalent in object detection or instance segmentation annotations, which ultimately result in the label convergence threshold. A detailed classification of these variations is illustrated in Figure 2, which includes five main categories:

1. **Imperfect Boundary Outlining (Bad Boundary):** This variation reflects a mismatch between the actual bounding box or mask and its annotated counterpart. Minor deviations are almost inevitable in all annotations and can be considered negligible.
2. **Mismatched Class Assignment (Wrong Class):** This occurs when the annotation assigns an incorrect class to an object, despite accurate localization.
3. **Overlooked Instances (Missed Instances):** Instances that are present in the image but not in the annotation are overlooked instances. In model evaluation, this would be similar to a false negative.

4. **Additional Instance:** Annotated instances without a corresponding object in the image, or for classes not represented in the dataset. This error is equivalent to a false positive in the model evaluation.
5. **Merged or Unmerged Instances:** Discrepancies in annotator decisions about whether to merge or separate instances, often influenced by occlusion or other factors, resulting in unnecessary merging or splitting of instances.

Our study contrasts with three other studies that simulate label variation. In Table 2, we compare the variations analyzed in both our study and theirs, demonstrating the comprehensive nature of our approach in examining real dataset variations across repeated labels in contrast to other studies that rely on synthetic label variations.

	Chan <i>et al.</i> [7]	Schubert <i>et al.</i> [34]	Chachula <i>et al.</i> [5]	Ours
Bad Boundaries (Type 1)	✓	✓	✓	✓
Wrong Class (Type 2)	✓	✓	✓	✓
Missed Instances (Type 3)	✓	✓	✓	✓
Additional Inst. (Type 4)	×	✓	✓	×
Merging Issue (Type 5)	×	×	×	✓
Redundancy (no error)	✓	×	×	×

Table 2. Comparative analysis of different approaches on handling specific annotation variation types.

Chan *et al.* [7] explore the impact of annotation variation on model performance and identify five types of variation, four of which overlap with our classification in Figure 2. They introduce specific variations such as mislabeled class/superclass and redundant annotations—the latter of which is not considered in our variation taxonomy, assuming it can be corrected by simple post-processing such as non-maxima suppression.

Chachula *et al.* [5] propose an algorithm for assessing and improving the quality of labels in object detection datasets. Their method, which is capable of flagging and re-annotating suspicious examples, covers all the variation types depicted in our analysis (Figure 2). This approach could potentially complement our methodology to more accurately identify annotation variations.

Schubert *et al.* [34] present an innovative technique for detecting label variation by simulating four types of variation outlined in our study and shown in Figure 2. Their method shows promise in identifying real label variation, suggesting the potential for integration with our findings to thoroughly address and correct variation in existing datasets. This approach could also potentially be added to our own methodology for identifying annotation variation.

However, the relevance of these studies to measuring convergence thresholds for real datasets with real label variation using simulated noise patterns remains uncertain, and

the effectiveness of their variation detection methods would require extensive manual verification.

3. Analyzed Dataset

In this section, we briefly introduce the datasets used in our study, which are summarized in Table 1. A crucial requirement for these datasets is that each image must be annotated by at least two independent raters. It is not mandatory for the raters to mark instances; if they conclude that no instance of the defined class set for the dataset is present in an image, it is acceptable to leave it unannotated. Based on these strict criteria, we selected segments from three well-known datasets in object detection: LVIS [11], COCO [19], and Open Images [16].

For LVIS, we use the doubly annotated subset of 5,000 images that was originally compiled to assess annotation consistency for v0.5 of the dataset. To ensure comparability with current SOTA models, we adjusted the data to align with v1.0 of the dataset, reducing the number of categories from 1723 to 1203 and excluding annotations of the removed categories. This subset serves as our primary benchmark due to its high quality, community recognition, consistent annotation guidelines, and complexity due to its wide class variety and long-tailed distribution. Unfortunately, this dataset lacks rater identification information, which divides the dataset into two generic subgroups.

Additionally, we examined two reannotated datasets introduced by Ma *et al.* [21], combined with the original COCO and Open Images datasets, to assess label convergence. However, these datasets have limitations that affect their validity and generalizability, detailed in Appendix A, which prevent us from determining label convergence for these datasets.

Lastly, we consider two smaller domain-specific datasets, VinDR-CXR [27] and TexBiG [39], which have the advantage of multiple raters per image and consistent labeling conventions throughout the annotation process, including rater identification metadata. While their smaller size and specific focus limit their impact compared to the results of the LVIS dataset, they technically provide the most suitable data for our analysis.

4. Determining Label Convergence

Determining label convergence is divided into two parts: (1) for cases where each image is annotated by exactly two independent raters, and (2) for cases where images are annotated by any number of raters. A key aspect in defining label convergence is to make the convergence threshold comparable to standard evaluation metrics, allowing for direct comparison with current model performance. To achieve this, we align label convergence with the mAP used in datasets such as COCO, LVIS, and PASCAL VOC. To

do this, we modify the regular mAP to work with different ground truth approximations instead of model predictions and a single ground truth. Since this still only allows us to evaluate between two annotators, we use Krippendorff’s Alpha version for object detection [39], which we also extend to work on instance segmentation, to assess the consistency of multiple annotators. By correlating these two values, we can easily but effectively determine the convergence for any number of annotators.

4.1. Two Annotators per Image

To adapt the mAP metric for evaluating label convergence between two annotators, we make specific modifications. Normally, mAP requires model confidence scores and a single ground truth, but for human annotations, we address these differences as follows:

- We set the confidence score of all annotations to 0.99, as human annotations are unlikely to produce severely overlapping detections of the same class. Unlike models, which then need to remove these excess detections through post-processing methods like non-maximum suppression.
- We randomly switch which annotator is considered ground truth and which is considered prediction, repeating the evaluation several times to account for variability.

Despite these adjustments, the evaluation remains comparable to the original mAP method. Here are the steps to compute the modified mAP:

1. Detections are sorted from highest to lowest confidence score.
2. Each detection is matched to a ground truth instance by IoU overlap, marking it as positive if matched or negative if not.
3. Positives and negatives are added to the Precision-Recall (PR) curve.
4. The area under the PR curve is calculated to obtain the AP for a category and a specific IoU threshold.
5. The AP values are averaged across all categories to obtain the mAP@[IoU threshold].

To statistically estimate the convergence threshold, we use bootstrapping, sampling 1,000 subsets, each with 10% of the available images of the respective dataset and evaluate the modified mAP on each of these subsets. By evaluating the 95% confidence interval of the sampling distribution, we capture the variability of the data and infer the convergence threshold for the entire dataset. The results show

a convergence threshold interval between 62.64 and 67.52 mAP for the LVIS consistency subset, with Co-DETR’s performance near the upper end at 66.8 mAP (see Table 3 and Figure 1). Due to issues with the COCO Reannotated and Open Images Reannotated datasets, these results are not included in the determination of the convergence threshold (see Appendix A).

Considerations for Extrapolation:

- Label convergence was estimated for a subset of the dataset, as only some images contain repeated labels. Although similar variations are expected in the remaining data, this is an extrapolation.
- Domain specificity of different subsets may create a domain gap between training and testing data, potentially affecting the generalizability of the convergence threshold.
- Labeling conventions need to be aligned. If a dominant group of annotators follows a particular labeling convention, models may overfit to that convention. Ensuring that annotators have roughly equal contributions is crucial for accurate label convergence estimation (discussed further in Section 5). It is reasonable to assume that if such a dominant labeling convention exists within the dataset, then following that labeling convention would allow a model to exceed the convergence threshold.

4.2. Multiple Annotators per Image

Since modified mAP cannot be used for images annotated by three or more annotators, we use Krippendorff’s Alpha for Object Detection [39] to assess annotation quality. This method, parameterized with different IoU thresholds such as mAP, provides a $K-\alpha$ value representing reliability between -1 and +1. A detailed explanation of this method is available in Appendix B.

To determine the convergence threshold for multiple annotators, we follow these steps:

1. Use the samples from the previous bootstrapping (LVIS, COCO, Open Images) and obtain the $K-\alpha$ values for different thresholds [0.5, 0.95, 0.05] and their mean. Also evaluate the mAP for all missing IoU thresholds as well.
2. Perform a linear least-squares regression (Figure 3) with $K-\alpha$ as the independent variable and mAP as the dependent variable. The regression shows a Pearson correlation of $\rho = 0.92$ and an R-squared value $R^2 = 0.85$, indicating a good model fit. The resulting equation is:

$$mAP = 0.836 \cdot \alpha + 0.197 \quad (1)$$

Dataset	Task	Img.	Samp. size	Mean	Std	Min	Max	CI L	CI U	SOTA
LVIS Cons. Subset [11]	det	5,000	500	65.08	1.25	61.49	69.32	62.64	67.52	66.8
	segm			59.78	1.26	55.51	63.79	57.32	62.24	60.2
COCO Reann. [21]	det	80,067	8,007	28.39	1.55	25.42	31.13	25.35	31.43	/
Open Images Reann. [21]	det	4,773	477	20.55	1.64	15.75	27.35	17.33	23.75	/
TexBiG [39]	det	2,257	226	81.89	0.82	78.17	84.57	80.28	83.50	49.84
	segm			72.87	0.92	69.40	75.95	71.07	74.66	44.06
VinDr-CXR [27]	det	15,000	1,500	47.08	1.51	42.27	52.80	44.12	50.05	31.4

Table 3. Statistics from bootstrapping with 1,000 samples on the five datasets analyzed, using a sampling size of 10% of the total dataset. This includes the lower and upper 95% confidence intervals, which we consider to be the convergence threshold interval. For LVIS, TexBiG and VinDr-CXR we also provide the latest SOTA results.

- Apply bootstrapping to the TexBiG and VinDr-CXR datasets to obtain $K-\alpha@[0.5, 0.95, 0.05]$ for TexBiG and $K-\alpha@[0.4]$ for VinDr-CXR, since these IoU thresholds were used in their respective leaderboards.
- Use the regression formula to infer the mAP values from the $K-\alpha$ values. Calculate the mean, standard deviation, and 95% confidence interval to estimate the convergence threshold interval, as shown in Table 1.

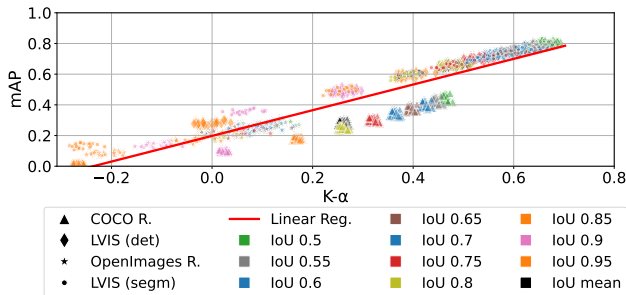


Figure 3. Best viewed digitally. Linear least-squares regression fit with $K-\alpha$ as the independent and mAP as the dependent variable. The scatter plots only show a fraction of the data points.

This method allows us to determine the convergence threshold for any number of annotators per image. However, these thresholds may be less reliable than directly evaluating mAP. While we use a simple linear model to describe the relationship between mAP and $K-\alpha$, future work with more data could explore better fitting models that do not assume linearity. We calculate the confidence interval empirically on the derived mAP values, as this better reflects the variability in the data than the error propagation from the regression model.

Although COCO Reannotated and Open Images Reannotated have issues (see Appendix A), they still represent real label variations. Thus, we use these datasets to express the relationship between mAP and $K-\alpha$ in our regression calculations, preferring real variations to synthetic ones.

5. Annotation Variation Type Analysis

After establishing the convergence threshold, we further analyze the distribution of variation types (Figure 4) and provide qualitative examples (Figure 5) of real label variations. To facilitate this, we use the FiftyOne [25] framework to visualize variation across the analyzed datasets. We introduce an algorithm that identifies different types of label variations and is designed to match as many examples as possible, although it tends to be strict in evaluating cases with three or more annotators. Detailed algorithmic descriptions, more qualitative examples, and an extended quantitative results section on the other datasets and instance segmentation can be found in Appendix C.

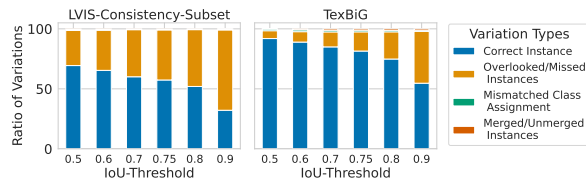


Figure 4. Quantitative results of the variation analysis on object detection, showing the ratio of correctly matched instances compared to unmatched instances. The most common variation types are 2 and 3 (Figure 2), which increase significantly as instances become harder to match due to higher IoU thresholds.

Figure 5 presents a representative example of the predominant issues, illustrating the four causes of human label variation and four of the five types of variation as shown in Figure 2. This brings us back to the issue of label convergence as discussed in Section 4.1. With this tangible example, we can see what happens when annotators follow different labeling conventions. If this happens on a large scale, where multiple annotators follow a similar labeling convention that is ambiguously described in the guideline, and this group has annotated a majority of the images, models may overfit to this specific convention, potentially exceeding the presumed convergence threshold. This can be avoided by

providing clear guidelines and ensuring that annotators have roughly equal contributions so that no single labeling convention dominates. Many other factors contribute to annotation variation, such as the selection of the annotation tool, which can affect perceived image quality.

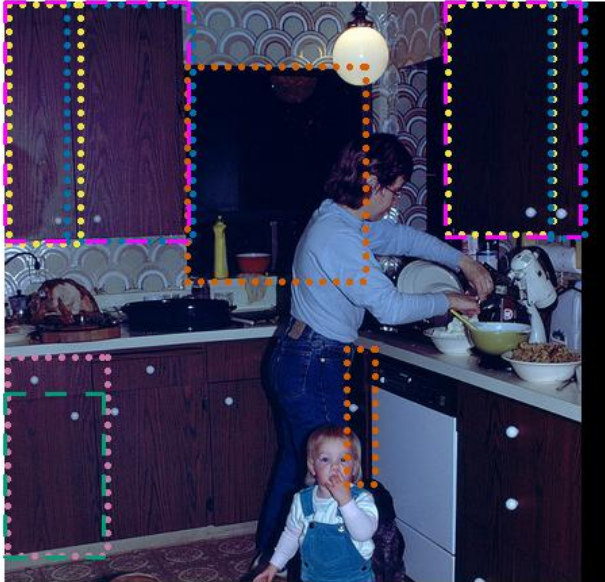


Figure 5. Example image from the LVIS Consistency Subset showing some of the variations. Only a selection of annotations from the cupboard class is visualized. The dotted line indicates coder A, while the dashed line indicates coder B. (a) At the top left and top right are two cases of a merging issue visible. This kind of inconsistencies should be prevented by a unambiguous guideline. While coder A separated the yellow and green areas of the two cupboards at the cupboard doors, coder B combined them as indicated by the magenta bounding boxes. (b) The two orange areas indicate an additional instance at the top, where a window is interpreted as a cupboard, while the bottom instance was found by coder A but missed by coder B. We attribute the first variation to image quality and the second one to an annotation error. (c) At the bottom left, another case of different labeling conventions is visible, where coder A covers the entire height of the cupboard, including the drawer, while coder B excludes the drawer from the cupboard. This could be due to either ambiguities in the guidelines or the skill of the annotator.

6. Conclusion

In our study, we address our central research question by enhancing the understanding of how annotation quality impacts model performance, introducing a straightforward and effective method for determining label convergence, which establishes a theoretical upper bound on model performance.

Our analysis shows that while state-of-the-art (SOTA) results approach this upper bound for the well-studied LVIS

dataset, the primary constraint is not model or data convergence but label quality. This suggests that models have sufficient capacity to handle the complexity of current object recognition problems. However, the current reliance on “gold standard” labels, despite inherent annotation variations, requires improvement in how problems are specified. We propose a combination of three key aspects to address this issue:

1. **Improving Annotation Quality:** Implement better guidelines and training for annotators to reduce variability and errors. The issue of label convergence should be addressed as early as possible, shifting the burden from the annotations themselves to clearer problem specification. This is particularly crucial for test data, where consistency is key to accurate evaluation.
2. **Including Multi-Annotated Data:** Since labeled data will likely always include some degree of variation, having a portion of the data annotated multiple times allows for an analysis of the extent of these variations. This enables models to become more robust to real-world scenarios by recognizing and accounting for annotation inconsistencies [2].
3. **Updating Evaluation Methods:** Reevaluate how strictly we distinguish between correct and incorrect annotations. Rather than aiming to eliminate all test set noise, we propose using the concept of label convergence as a measure of ambiguity for unavoidable annotation inconsistencies. This approach recognizes that some level of variability in labels is inevitable and should be incorporated into the evaluation process, leading to a more flexible and realistic assessment of model performance.

In summary, our study emphasizes the critical role of label quality in achieving optimal model performance, and we propose a more nuanced approach to problem specification and evaluation that takes into account unavoidable annotation variability.

Code for our FiftyOne plugin [25] can be accessed at <https://github.com/Madave94/multi-annotator-toolkit>.

7. Outlook

Our study is limited by focusing solely on computer vision, without comparisons to fields like natural language processing (NLP). We believe these domains differ significantly, as language inherently involves ambiguities, while computer vision often aims for a single, but sometimes unattainable, ground truth.

References

- [1] Cathaoir Agnew, Ciarán Eising, Patrick Denny, Anthony Scanlan, Pepijn Van De Ven, and Eoin M. Grua. Quantifying the Effects of Ground Truth Annotation Quality on Object Detection and Instance Segmentation Performance. *IEEE Access*, 2023. 2, 3
- [2] Abhishek Anand, Negar Mokhberian, Prathyusha Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. Don't Blame the Data, Blame the Model: Understanding Noise and Bias When Learning from Subjective Annotations. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertainNLP 2024)*, 2024. 8
- [3] Ali Borji and Seyed Mehdi Iranmanesh. Empirical Upper Bound in Object Detection and More, 2019. arXiv:1911.12451 [cs]. 3
- [4] Andreas Bär, Jonas Uhrig, Jeethesh Pai Umesh, Marius Cordts, and Tim Fingscheidt. A Novel Benchmark for Refinement of Noisy Localization Labels in Autolabeled Datasets for Object Detection. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference Workshops (CVPRW)*, 2023. 2, 3
- [5] Krystian Chachula, Jakub Łyskawa, Bartłomiej Olber, Piotr Frątczak, Adam Popowicz, and Krystian Radlak. Combating noisy labels in object detection datasets, 2023. arXiv:2211.13993 [cs]. 5
- [6] Simon Chadwick and Paul Newman. Training Object Detectors With Noisy Data, 2019. arXiv:1905.07202 [cs]. 3
- [7] Abraham Chan, Arpan Gujarati, Karthik Pattabiraman, and Sathish Gopalakrishnan. Evaluating the Effect of Common Annotation Faults on Object Detection Techniques. In *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, 2023. 5
- [8] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 1960. 4
- [9] DeepLearningAI. A Chat with Andrew on MLOps: From Model-centric to Data-centric AI, Mar. 2021. 1
- [10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88:303–338, 2010. 2
- [11] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2019. 2, 4, 5, 7
- [12] Khiem H. Le, Tuan V. Tran, Hieu H. Pham, Hieu T. Nguyen, Tung T. Le, and Ha Q. Nguyen. Learning From Multiple Expert Annotators for Enhancing Anomaly Detection in Medical Image Analysis. *IEEE Access*, 2023. 3
- [13] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 2007. 4
- [14] Yucheng Hu and Meina Song. Crowd R-CNN: An Object Detection Model Utilizing Crowdsourced Labels. In *Proceedings of the International Conference on Vision, Image and Signal Processing (ICVISIP)*, 2019. 3
- [15] Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011. 4
- [16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4. *International Journal of Computer Vision (IJCV)*, 2020. 5
- [17] Hengduo Li, Zuxuan Wu, Chen Zhu, Caiming Xiong, Richard Socher, and Larry S. Davis. Learning From Noisy Anchors for One-Stage Object Detection. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020. 3
- [18] Junnan Li, Caiming Xiong, Richard Socher, and Steven Hoi. Towards Noise-resistant Object Detection with Noisy Annotations, 2020. arXiv:2003.01285 [cs]. 3
- [19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 4, 5
- [20] Chengxin Liu, Kewei Wang, Hao Lu, Zhiguo Cao, and Ziming Zhang. Robust Object Detection with Inaccurate Bounding Boxes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3
- [21] Jiaxin Ma, Yoshitaka Ushiku, and Miori Sagara. The Effect of Improving Annotation Quality on Object Detection Datasets: A Preliminary Study. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference Workshops (CVPRW)*, 2022. 2, 4, 5, 7, 11
- [22] Jiafeng Mao, Qing Yu, and Kiyoharu Aizawa. Noisy Localization Annotation Refinement for Object Detection. *IEICE Transactions on Information and Systems*, 2021. 2, 3
- [23] Jiafeng Mao, Qing Yu, Yoko Yamakata, and Kiyoharu Aizawa. Noisy Annotation Refinement for Object Detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. 3
- [24] Elena Merdjanovska, Ansar Aynetdinov, and Alan Akbik. NoiseBench: Benchmarking the Impact of Real Label Noise on Named Entity Recognition. arXiv:2405.07609 [cs]. 3
- [25] B. E. Moore and J. J. Corso. Fiftyone, 2020. GitHub: <https://github.com/voxel51/fiftyone>. 7, 8
- [26] Joseph Nassar, Viveca Pavon-Harr, Marc Bosch, and Ian McCulloh. Assessing data quality of annotations with krippendorff alpha for applications in computer vision, 2019. arXiv:1912.10107 [cs]. 4
- [27] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 2022. 2, 3, 5, 7
- [28] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2
- [29] Junran Peng, Xingyuan Bu, Ming Sun, Zhaoxiang Zhang, Tieniu Tan, and Junjie Yan. Large-Scale Object Detection

- in the Wild From Imbalanced Multi-Labels. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020. [3](#)
- [30] Hieu H. Pham, Khiem H. Le, Tuan V. Tran, and Ha Q. Nguyen. Improving Object Detection in Medical Image Analysis through Multiple Expert Annotators: An Empirical Investigation, 2023. arXiv:2303.16507 [cs]. [3](#)
- [31] Hieu H. Pham, Ha Q. Nguyen, Hieu T. Nguyen, Linh T. Le, and Lam Khanh. An Accurate and Explainable Deep Learning System Improves Interobserver Agreement in the Interpretation of Chest Radiograph. *IEEE Access*, 10, 2022. [3](#)
- [32] Barbara Plank. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022. [2](#), [3](#), [4](#)
- [33] V Ribeiro, S Avila, and E Valle. Handling inter-annotator agreement for automated skin lesion segmentation. arxiv, 2019. arXiv:1906.02415 [cs]. [4](#)
- [34] Marius Schubert, Tobias Riedlinger, Karsten Kahl, Daniel Kröll, Sebastian Schoenen, Siniša Šegvić, and Matthias Rottmann. Identifying label errors in object detection datasets by loss inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision WACV*, 2024. [2](#), [5](#)
- [35] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. [2](#)
- [36] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE/CVF International Conference in Computer Vision (ICCV)*, 2017. [2](#)
- [37] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning From Noisy Labels by Regularized Estimation of Annotator Confusion. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2019. [2](#)
- [38] David Tschirschwitz, Christian Benz, Morris Florek, Henrik Norderhus, Benno Stein, and Volker Rodehorst. Drawing the Same Bounding Box Twice? Coping Noisy Annotations in Object Detection with Repeated Labels. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 2023. [2](#), [3](#), [11](#)
- [39] David Tschirschwitz, Franziska Klemstein, Benno Stein, and Volker Rodehorst. A Dataset for Analyzing Complex Document Layouts in the Digital Humanities and its Evaluation with Krippendorff’s Alpha. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 2022. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [12](#)
- [40] Mengmeng Xu, Yancheng Bai, and Bernard Ghanem. Missing Labels in Object Detection. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference Workshops (CVPRW)*, 2019. [2](#)
- [41] Zizhao Zhang, Han Zhang, Sercan O. Arik, Honglak Lee, and Tomas Pfister. Distilling Effective Supervision From Severe Label Noise. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020. [2](#)
- [42] Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with Collaborative Hybrid Assignments Training. In *Proceedings of the IEEE/CVF International Conference in Computer Vision (ICCV)*, 2023. [1](#)