IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 August 2024, accepted 21 September 2024, date of publication 24 September 2024, date of current version 7 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3467008

RESEARCH ARTICLE

Annotation Quality Versus Quantity for Object Detection and Instance Segmentation

CATHAOIR AGNEW[1,2,3], ANTHONY SCANLAN[1,2,3], PATRICK DENNY[3,4], (Member, IEEE), EOIN M. GRUA[1,2,3], PEPIJN VAN DE VEN[1,2,3], AND CIARÁN EISING[1,2,3], (Senior Member, IEEE)

[1]Department of Electronic and Computer Engineering, University of Limerick, Limerick, V94 T9PX Ireland
[2]CONFIRM Centre for Smart Manufacturing, University of Limerick, Limerick, V94 T9PX Ireland
[3]Data-Driven Intelligent Computer Engineering (D[2]iCE) Research Group, University of Limerick, Limerick, V94 T9PX Ireland
[4]Department of Computer Science and Information Systems (CSIS), University of Limerick, Limerick, V94 T9PX Ireland

Corresponding author: Cathaoir Agnew (cathaoir.agnew@ul.ie)

This work was supported by the Science Foundation Ireland (SFI) under Grant 16/RC/3918 (CONFIRM Centre).

ABSTRACT Deep learning-based computer vision models are typically data-hungry, resulting in the rise of dataset sizes. The consensus for computer vision datasets is that larger datasets lead to better model performance. However, the quality of the datasets is often not considered. Annotating datasets for fully supervised object detection and instance segmentation tasks requires a significant investment in time, effort, and cost. In practice, due to the large sample sizes needed, this often leads to inaccuracies in the annotation process. This research aims to understand and quantify the impact of annotation quality and quantity on the performance of object detection and instance segmentation models. Specifically, the research aims to investigate how introducing additional data with varying levels of annotation quality affects mean average precision (mAP) performance. To investigate the relationship between annotation quality and quantity, subsets of the COCO and ADE20K datasets are used. For each of the datasets, three types of annotation uncertainty are added to the annotations, which are localization uncertainty, incorrect class labels, and missing annotations. Mask-RCNN, YOLACT, and Mask2Former models are trained on a variety of sample sizes for varying levels of annotation uncertainties. The results indicate there is utility in adding additional data of lesser annotation quality. The extent of the benefits of the additional data is directly related to how degraded the annotations' are. Furthermore, the results show that all three annotation uncertainties negatively affect mAP performance, with incorrect class labels degrading mAP performance the most, followed by missing annotations and lastly localization uncertainty.

INDEX TERMS Annotation uncertainty, computer vision, instance segmentation, object detection, supervised learning.

I. INTRODUCTION

The size of datasets used for training computer vision models is steadily rising, influenced by the prominence of data-hungry deep learning-based architectures [1]. This is illustrated by a dataset size of 11,540 images for the Pascal Visual Object Classes (VOC) dataset [2] published in 2012, 164,000 images for the Common Objects in Context (COCO) dataset [3] in 2014 and finally 2 million images for the Objects365 dataset released in 2019 [4]. The Segment Anything 1 Billion (SA-1B) dataset released in 2023 [5] contains 11 million images. However, class labels are not provided as the dataset uses class-agnostic mask annotations. The use of larger datasets provides models with more diverse training examples which in turn result in better representation learning [6] along with reducing the effect of overfitting [7].

On large-scale benchmark datasets for object detection and instance segmentation, fully supervised approaches have produced impressive results [8], [9]. The performance of fully-supervised object detection and instance segmentation

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea Bottino.

models relies on the dataset used to train the model [10], including the annotations that are utilized for the ground truth. For object detection, the ground truth annotation consists of a bounding box, which is used to locate the object in the image along with a class label to identify the object category. Polygon masks are used in place of bounding boxes to outline the boundary of an object for instance segmentation. The process of annotating datasets for supervised learning tasks in computer vision can be time-consuming and reaching dataset sizes of the scale of COCO requires significant investment in both time and money. To illustrate, annotation of the COCO dataset took approximately 60,000 worker hours, with an estimated 79.2 seconds per polygon mask [3]. The median time for each annotation within the ImageNet Visual Recognition dataset [11] was 42.4 seconds to complete when utilizing a crowd-sourcing approach that has been specifically designed for bounding box annotation purposes [12].

Research has shown that, as the quantity of the training dataset grows, performance on mAP [13] for computer vision tasks increases logarithmically [6]. This has motivated the rise in dataset sizes along with methods to increase the size and diversity of the training dataset, such as data augmentation [14], semi-supervised learning [15], and generating synthetic data using generative adversarial networks (GANs) [16]. Foundation models, which often utilize self-supervised and semi-supervised learning techniques on large quantities of data, have shown improvements in the pretext task in computer vision. Utilizing a vision foundation model, Wang et al. set a new COCO test-dev benchmark mean average precision (mAP) of 0.654 [9]. Whilst there have been improvements with foundation models in computer vision tasks, there is still the need for annotations for the downstream tasks.

The literature has established the significance of high-quality ground truth annotations for computer vision tasks [10]. Research has also been conducted on methods attempting to identify and correct noisy labels for both image classification [17], [18], [19], [20], [21] and object detection [22], [23], [24]. However, to the authors' knowledge, there is limited research attempting to investigate the tradeoff in mAP performance between annotation quality and quantity for object detection and instance segmentation.

The main contribution of this paper is the investigation of the relationship between annotation quality and quantity for object detection and instance segmentation on mAP performance. This is accomplished by creating subsets of the COCO and ADE20K datasets and introducing annotation uncertainty into a selection of the ground truth annotations to varying degrees of severity. The main contributions are summarised as follows:

- An investigation into the relationship between annotation quality and quantity on mAP performance is undertaken for both object detection and instance segmentation.
- This research investigates and quantifies the effects of introducing additional data of lesser annotation quality to the training dataset.
- The effect of localization uncertainty, incorrect class labels, and missing annotations are quantified on mAP performance.

The paper is structured as follows. An overview of the related work is presented in Section II. In Section III, an explanation of the annotation uncertainty used is given. This is followed by a description of the experiment in Section IV. Then, in Section V, a presentation of the results is given, with the results being analyzed and discussed in Section VI. Finally, Section VII summarizes this work's conclusions.

II. RELATED WORK

Sun et al. investigated the impact of data on deep learning methodology [6]. As part of this research, the effect of pre-training sample size and performance was studied for object detection and semantic segmentation. The authors built upon the datasets used in [25] and [26], and created the JFT-300M dataset. This dataset consists of 300M images with 375M labels, containing 18,291 categories. The authors estimated there is approximately 20% category label noise in the JFT-300M dataset as the labels were automatically generated using an algorithm that made use of raw web signals, user feedback, and connections between web pages. Faster-RCNN models [27] were trained on three randomly selected subsets of the JFT-300M of size 10M, 30M, and 100M images. The weights from the three trained models were then used to initialize the model weights for training on the COCO and Pascal VOC datasets. It was found that as the pre-training dataset increases, performance increases logarithmically for both object detection and semantic segmentation. Whilst this research investigated the effects of pre-training sample size, the results do not directly extend to the effects of the training set size for the downstream task.

Shahinfar et al. investigated the relationship between sample size and per-class performance for autonomous wildlife monitoring [28]. To explore the effect of the number of training images and the performance of the image classification models, the authors created 7 subsets of the training dataset containing 10, 20, 50, 150, 500, and 1000 images per class. The authors trained 6 image classification models as part of their study using the ResNet architecture [29] with 18, 50, and 152 layers along with the DenseNet architecture [30] with 121, 161, and 201 layers. The experiments were conducted for each of the datasets collected from Africa, Australia, and North America, and the findings were consistent across each of the geographical datasets. The authors concluded there was a logarithmic relationship between the number of training images and model performance, along with a logarithmic relationship between the false positive rate and the inverse of the number of training images. Whilst the study provides insight into the effects of training sample size for image classification,

this research does not investigate the effects of annotation uncertainty.

In our previous work [10], the effect of localization uncertainty was quantified for various levels and types of induced noise to investigate the relationship between localization uncertainty and mAP performance for object detection and instance segmentation. A subset of the COCO dataset along with the Cityscapes dataset [31] was used to investigate the relationship between localization uncertainty and mAP performance. For the COCO dataset, a strong linear relationship was found between both noise types and mAP performance. When investigating the per-class performance for both object detection and instance segmentation the degradation across classes varied, suggesting that localization annotation quality and mAP performance is class-dependent. Whilst this work highlights the importance of localization annotation quality, the entire datasets were induced with annotation uncertainty. These results would not fully extend to the relationship between annotation quality and quantity. Furthermore, the effects of incorrect class labels and missing annotations were not considered in this work.

Whilst there exist works in the literature that investigate the effects of annotation uncertainty for image classification [17], [18], [19], [20], [21], object detection [22], [23], [24], [32], [33], instance segmentation [10], [34] and semantic segmentation [35], the relationship between annotation quality and quantity is relatively unexplored. The aim of this study is to examine the relationship between the quality and quantity of annotations and their influence on mAP performance for object detection and instance segmentation. Furthermore, this research aims to analyze the effects of incorporating additional data with lower annotation quality into the training dataset and its subsequent impact on mAP performance.

III. ANNOTATION UNCERTAINTY

This work investigates the effect of different aspects of annotation uncertainty as outlined in this section. An example for each annotation type and the degradation used are shown in Fig. 1.

A. LOCALIZATION UNCERTAINTY

The localization of the annotation, that is, the coordinates of the bounding boxes and polygon masks, will have uncertainty introduced following the same algorithms as set out in our previous research [10]. Gaussian radial noise was introduced to each vertex of the polygon masks and bounding boxes to simulate annotation uncertainty. For the bounding boxes, this was introduced following Equation (1) from [10]. In Equation (1) the terms w and h represent the width and height of the bounding box respectively while x and y correspond to the most upper left-hand point following the COCO bounding box format. σ is the standard deviation used for the normally distributed ($\mathcal{N}$) noise used to degrade the annotation with x_r, y_r, w_r, h_r representing the new datapoints with radial noise for the bounding box. For x_r, y_r, the normally distributed noise has a standard deviation of 1, as this was found to better represent localization uncertainty observed in datasets. As for the polygon masks, the full method for adding radial noise is described in [10]. For the experiments, integer values of 2 and 5 were used to replicate a small and moderate amount of noise. Fig. 1 (d) can be considered in the context of our previous work [10], which shows the impact of localization uncertainty on mAP performance is considerable. For a one-unit increase in σ, we predict a decrease of 0.0241 and 0.0135 for object detection and instance segmentation respectively on mAP performance when expressed as a decimal. As seen in Fig. 2, when $\sigma = 5$, the model struggles to detect small objects.

$$\begin{aligned} x_r &= x - |\mathcal{N}(0, 1^2)| \\ y_r &= y - |\mathcal{N}(0, 1^2)| \\ w_r &= w + |\mathcal{N}(0, \sigma^2)| \\ h_r &= h + |\mathcal{N}(0, \sigma^2)| \end{aligned} \tag{1}$$

B. INCORRECT CLASS LABELS

An investigation into the effect of incorrect class labels is undertaken for this work. We consider incorrect class labels when the category label given for the annotation is not the true label for the object. This is shown in Fig. 1 (f) when one of the elephants is given the class label spoon. The COCO detection dataset consists of a total of 80 classes. For each class label, an integer value is assigned to map the class label to the integer representation. For each annotation, a uniformly distributed float between the ranges of 0-1 inclusive is generated. If the float is below a defined threshold value, the integer representing the class label for the given annotation was randomly sampled from a list of the class label integers, ensuring the correct class label was not selected. This results in an approximate amount of incorrect class labels controlled by the threshold value chosen. To investigate the impact of incorrect class labels, four levels of class label noise were used: 25%, 50%, and 75%.

C. MISSING ANNOTATIONS

The effect of missing annotations for objects of interest in the dataset was investigated. We consider a missing annotation when an object of interest in an image is not annotated. This is shown in Fig. 1 (e), where only one of the elephants is annotated. Following the same randomization process used to generate incorrect class labels, annotations were randomly selected to be omitted from the dataset with three levels of approximate missing annotations used: 25%, 50%, and 75%. A breakdown of the number of annotations per missing annotations dataset is given in Table 3. This informs and allows us to compare the number of annotations present for each of the missing percentage thresholds along with each of the datasets' sizes.

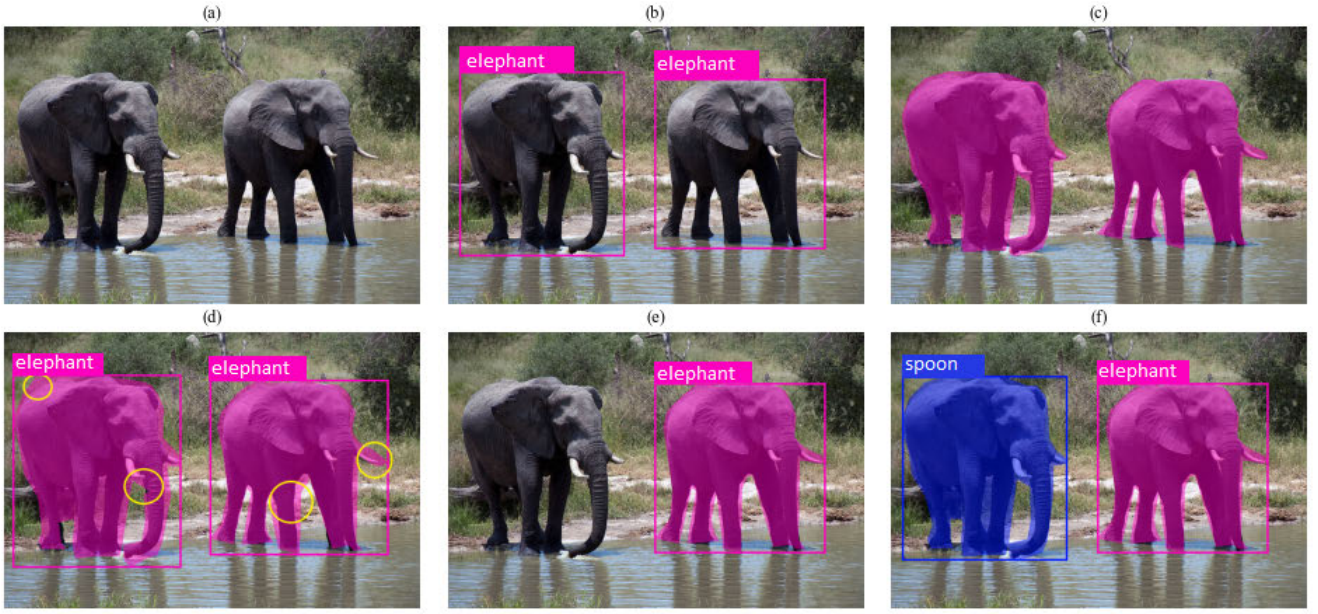

FIGURE 1. Example of annotations. (a) Original Image, (b) Bounding Boxes, (c) Polygon Masks, (d) Radial Noise $\sigma = 5$, (e) Missing Annotation, (f) Incorrect Class Label. Image from the COCO Dataset [3].

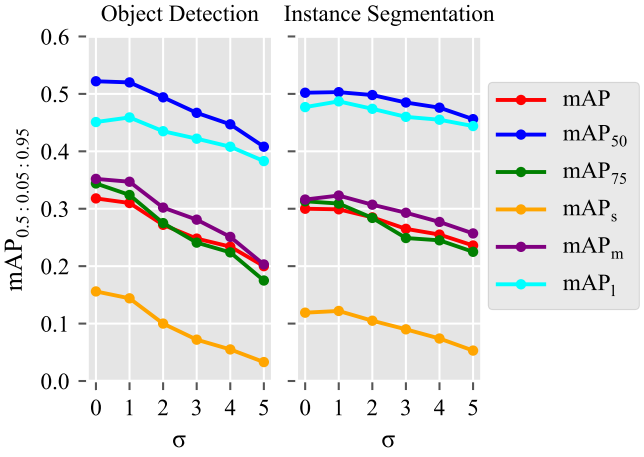

FIGURE 2. Results of the effect of localization uncertainty on the COCO Dataset, from our previous research [10].

IV. EXPERIMENTAL DESIGN

A. DATASET

A selection of subsets of the COCO 2017 detection dataset [3] was created for the experiments in this work. The COCO 2017 detection training dataset has approximately 118,000 images. Four datasets of randomly selected images of sizes 10,000, 25,000, 50,000, and 100,000 from the COCO training dataset were used for the investigation into the relationship between annotation quality and quantity. The subsets remain fixed throughout each of the experiments, as they are created as individual datasets. The sample sizes were chosen to investigate a 2.5x, 5x, and 10x relationship relative to the baseline dataset size of 10,000. Starting with 100,000 images, each subset is used to generate the next. For example, in the 50,000 images dataset, these images were randomly selected from the 100,000 images dataset. This helps ensure a fairer comparison when investigating the relationship between sample size and mAP performance, as the datasets are dependent and contain an overlap of shared images.

These subsets will have induced annotation uncertainty as described in Section III. We explore two scenarios in this work. Firstly, a percentage of annotations are randomly selected to have annotation uncertainty introduced. This results in a random selection of annotations having annotation uncertainty introduced rather than each of the annotations in the image. In this scenario, we may have one annotation of a class induced with annotation uncertainty while the remaining annotations of the same or other classes remain untouched for the given image.

Secondly, we investigate the effect of merging a clean dataset, that is, the original ground truth annotations, and a second dataset of lesser annotation quality to create a new dataset. This results in a selection of images whose annotations all have annotation uncertainty introduced within the dataset, whilst the remaining images remain with the ground truth annotations. To create the clean and second dataset of lesser quality, the individual subsets are further split into two datasets, with the split ratio dependent on the percentage of degradation being introduced. The first split will be the clean dataset, whose annotations will remain untouched. The second dataset will introduce annotation uncertainties. This ensures the clean dataset and the second dataset of lesser quality are independent of one another. Finally, these two datasets, the clean dataset and the second dataset of lesser quality, are then merged to create the final dataset. The ADE20K dataset [36] was also used to

provide results for a reduced experiment set to lessen the computational workload. The ADE20K dataset has finer scale annotations in comparison to COCO, as ADE20K is a semantic segmentation dataset with pixel-wise labels. The same process as outlined for the COCO dataset was used for creating the subsets of the ADE20k, with the subsets being of size 2,000, 5,000, 10,000, and 20,000. The original validation dataset of 2,000 images was split into 1,020 images used for validation and 980 images used as the hold-out test dataset for the experiments. As for the COCO experiments, the original COCO validation and test-dev datasets were used.

B. TRAINING SETUP

Mask-RCNN [37], YOLACT [38], and Mask2Former [39] were the model architectures chosen for the experiments due to their abilities to predict both object detection and instance segmentation outputs. The Mask-RCNN architecture is a two-stage convolutional neural network (CNN) detector, whereas the YOLACT architecture is a single-stage CNN detector. Mask2Former, on the other hand, is a transformer-based detector. The MMdetection framework [40] was used to train each of the models for the experiments. Unless otherwise stated, all models utilized a ResNet-50 [29] backbone, pre-trained on Imagenet [41], with the remainder of the model weights randomly initialized. All experiments used the same training parameters for the respective models. Mask-RCNN models were trained for 12 epochs with an effective batch size of 32. A stochastic gradient descent (SGD) optimizer with a learning rate of 0.02, a weight decay of 0.0001, and a momentum of 0.9 was utilized. At epochs 8 and 11, a learning rate scheduler was utilized to drop the learning rate by a factor of 10. Mask-RCNN models that utilized a Swin-T [42] or ConvNeXt-T [43] backbone were trained for 12 epochs with a batch size of 16. An AdamW optimizer was utilized with a learning rate of 0.0001 and a weight decay of 0.05. The YOLACT models trained for 55 epochs with an effective batch size of 32, utilizing an SGD optimizer with a learning rate of 0.001, a weight decay of 0.0005, and a momentum of 0.9, with a learning rate scheduler dropping the learning rate by a factor of 10 at epochs 20, 42, 49 and 52.

The Mask2Former models were trained for 50 epochs, utilizing an AdamW optimizer with a learning rate of 0.00005 and a weight decay of 0.05. Early stopping was implemented for the Mask2Former models, with a patience of 10 epochs monitoring the validation mAP score in decimals, with a minimum change of 0.005 in the mAP score required to continue the training. A learning rate scheduler was not used since the Mask2Former models used early stopping. If a learning rate scheduler had been utilized, certain models could have converged before reaching the epoch number for which the learning rate scheduler would have reduced the learning rate, thus resulting in different training parameters. Maintaining a consistent learning rate guaranteed fair comparisons across all experiments involving Mask2Former models. All experiments used distributed training between two NVIDIA A100 40Gb GPUs. The Mask-RCNN models were used to investigate the effect of merging a clean and second dataset of lesser quality on the COCO dataset; however, for computational workload reasons, YOLACT and Mask2Former were not used for these experiments. Lastly, again for reasons of computational load, only Mask-RCNN models were trained on the ADE20K dataset.

V. RESULTS

The results from the experiments were obtained from the official test-dev set of the COCO dataset and the created hold-out test dataset for ADE20k. COCO's primary metric $mAP_{0.50:0.05:0.95}$ (mAP) [13] is the metric of interest for the experiments and is expressed in decimal format. The experiments are divided into 3 subsections: localization uncertainty, incorrect class labels, and missing annotations. A second metric, τ, is calculated by normalizing each of the degraded models' test scores to the model trained on the original ground truth annotations for each given sample size. This gives insight into the relationship between annotation degradation and sample size. τ can be interpreted as the percentage remaining of the chosen reference mAP score. A comparison across the annotation error types and τ is shown in Table 2, with the scores being averaged across all three model architectures and each of the four dataset sizes: 10,000, 25,000, 50,000, and 100,000. Finally, the inference results for varying Mask2Former models trained on 100k images can be seen in Fig. 22 to understand model performance further.

A. LOCALIZATION UNCERTAINTY

A total of 136 (72 for Mask-RCNN, 32 for YOLACT, and 32 for Mask2Former) models were trained to investigate the relationship between the training sample size and the effect of localization uncertainty on mAP performance. The results on the COCO dataset can be seen in Fig. 3. The legend for the respective figures can be interpreted as follows: 100% Clean refers to the original ground truth annotations being used, 75% Clean 25% $\sigma = 5$ refers to 75% of the original ground truth annotations being used with the remaining 25% of the annotations having localization uncertainty introduced with a $\sigma = 5$. Lastly, 100% $\sigma = 5$ refers to all the annotations having localization uncertainty introduced with a σ value = 5. The comparison between the randomly selected percentage of annotations degraded and merging a clean and a second dataset of lesser quality is seen in Fig. 19. Finally, the results on the ADE20K datasets can be seen for both object detection and instance segmentation in Fig. 4. The results suggest that perfectly labeled bounding boxes and polygon masks yield the highest scores per sample size, nonetheless, some minor levels of localization uncertainty do not have a significant detrimental impact on model performance.

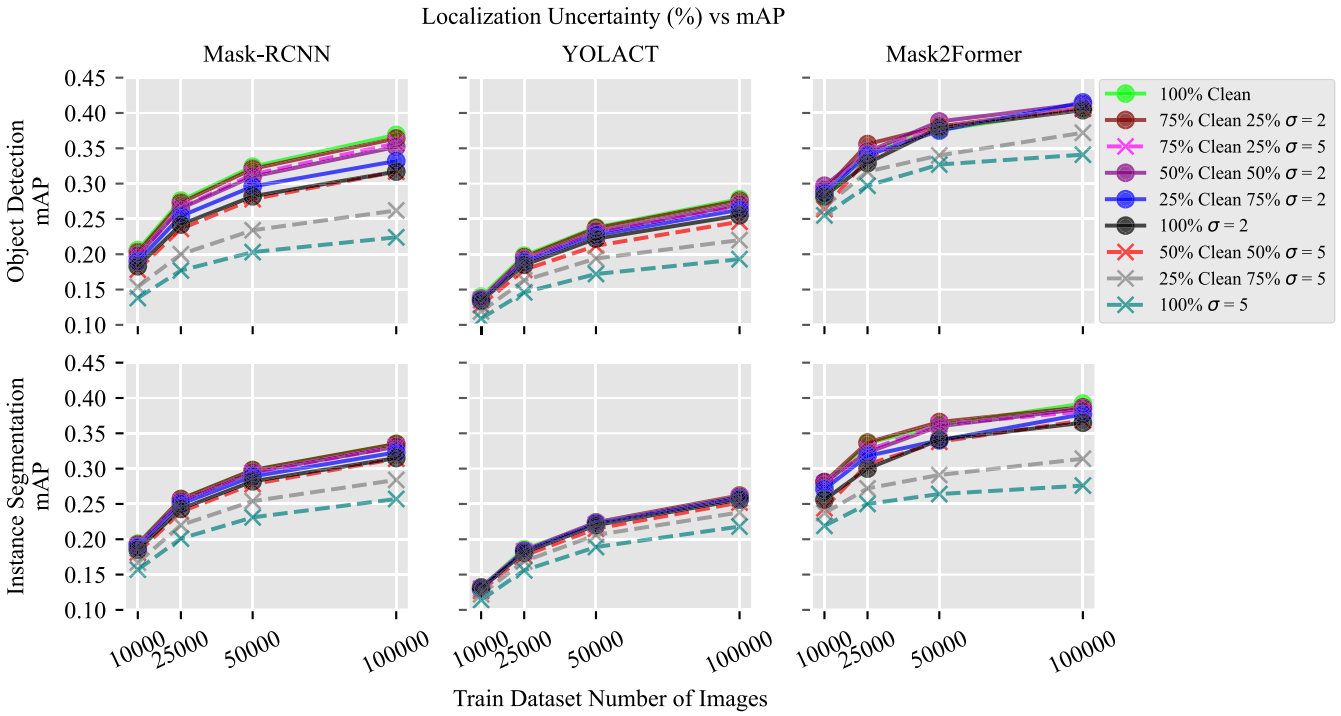

FIGURE 3. COCO results for varying localization uncertainty (σ is the standard deviation used for generating Gaussian radial noise) on mAP Performance.

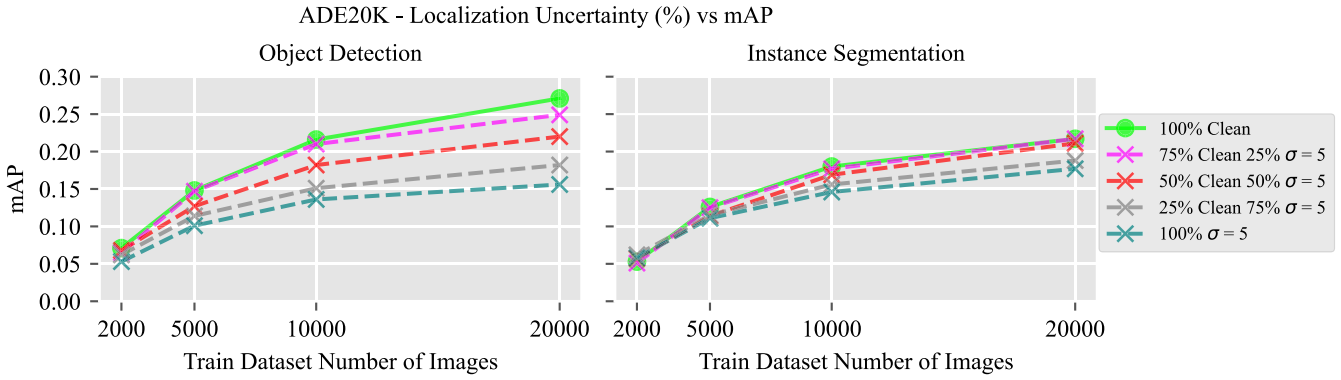

FIGURE 4. Mask-RCNN model results on ADE20K subsets for varying localization uncertainty on mAP performance.

B. INCORRECT CLASS LABELS

A total of 96 (72 for Mask-RCNN, 12 for YOLACT, and 12 for Mask2Former) models were trained to investigate the relationship between the training sample size and the effect of incorrect class labels on mAP performance. The results on the COCO dataset are seen in Fig. 5. The legend for the respective figures can be interpreted as follows: 100% Clean refers to the original ground truth annotations being used, 25% incorrect class labels refer to 25% of the annotations with induced class labels errors. Due to the JSON submission file being too large for the evaluation server to handle, the validation results are used for the instance segmentation results for Mask2Former in this section. The results for merging a clean and a second dataset of lesser quality are seen in Fig. 17, with direct comparisons seen in Fig. 20. Finally, the results on the ADE20K datasets can be seen in Fig. 6. The results from this experiment indicate that annotating class labels requires careful consideration as it can severely impact mAP performance.

C. MISSING ANNOTATIONS

A total of 84 (60 for Mask-RCNN, 12 for YOLACT, and 12 for Mask2Former) models were trained to investigate the relationship between the training sample size and the effect of missing annotations on mAP performance. The results on the COCO dataset are seen in Fig. 7. The legend for the respective figures can be interpreted as follows: 100% Clean refers to the original ground truth annotations being used, 25% missing annotations refer to 25% of the annotations being dropped from the dataset. The results between the randomly selected

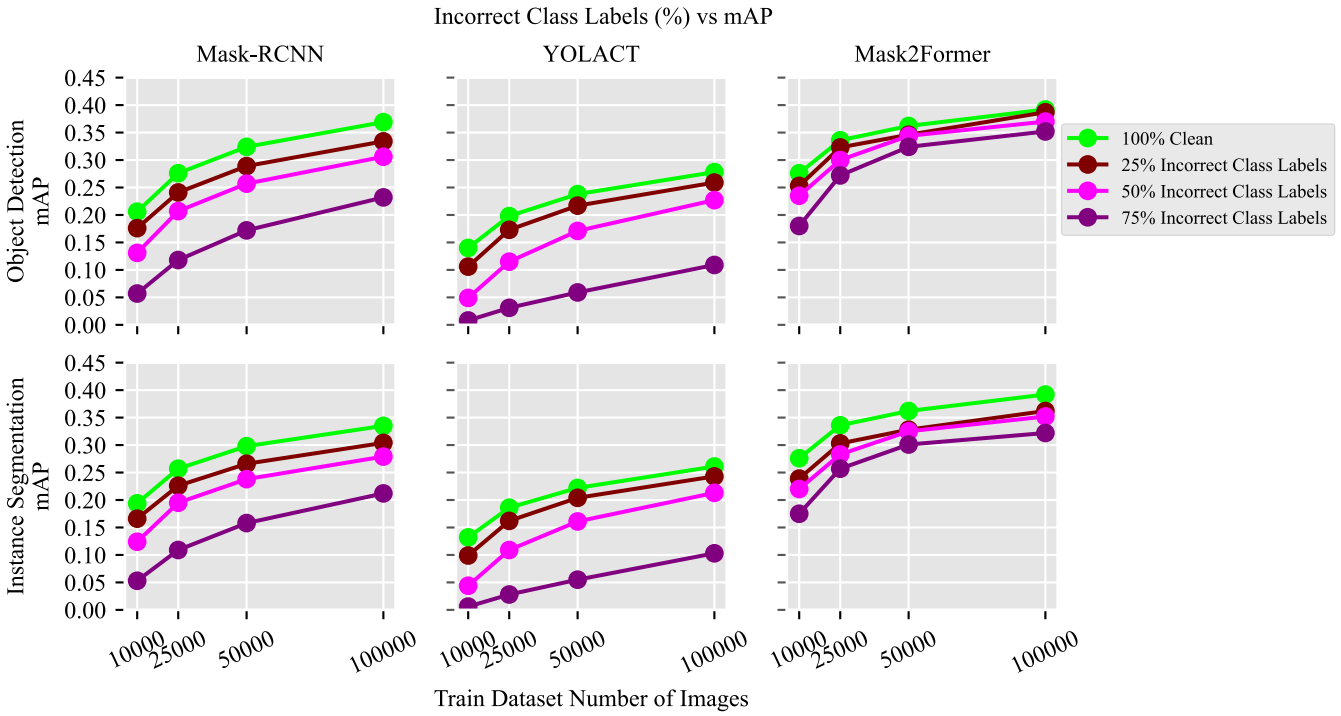

FIGURE 5. COCO results for varying incorrect class labels on mAP performance.

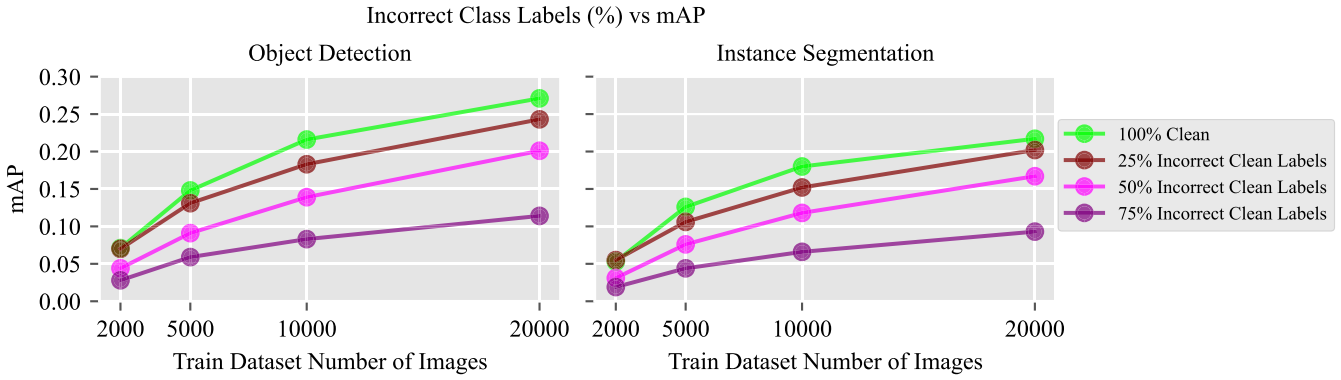

FIGURE 6. Mask-RCNN model results on ADE20K subsets for varying incorrect class labels on mAP performance.

percentage of annotations degraded and merging a clean and a second dataset of lesser quality is seen in Fig. 18. Finally, the results on the ADE20K datasets can be seen in Fig. 8. The findings from this experiment indicate the quantity of annotations is more important than the quantity of images in the dataset. This demonstrates the importance of annotating every object of interest in the dataset before gathering further data.

D. EFFECT OF BACKBONE

A total of 80 Mask-RCNN models were trained to investigate the relationship of the backbone used on each annotation uncertainty. Three backbones were considered for this work, ResNet-50, Swin-T, and ConvNeXt-T, which focused on Mask-RCNN models and the COCO dataset to reduce the computational workload. The backbones have approximately 26M, 28M, and 28M parameters, respectively. The results of the localization uncertainty on the COCO dataset can be seen in Fig. 9. The results of the incorrect class labels can be seen in Fig. 10. Finally, the results for the missing annotations can be seen in Fig. 11. The results suggest that regardless of the backbone used, the trends are still apparent across each of the annotation uncertainties.

E. COMBINED ANNOTATION UNCERTAINTIES

A total of 36 (24 for Mask-RCNN, and 12 for YOLACT) models were trained to investigate the relationship between the training sample size and the effect of combining the annotation uncertainties on mAP performance. The combined annotation uncertainties were created as follows.

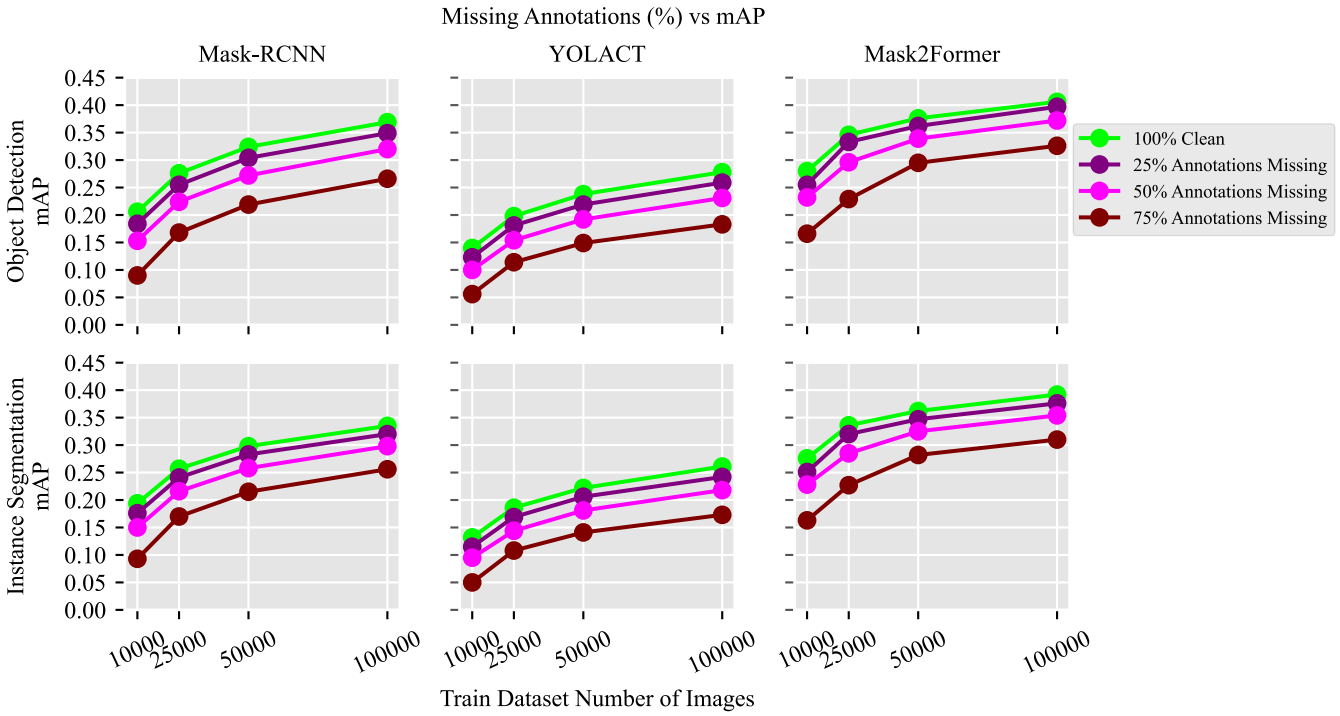

FIGURE 7. COCO results for varying missing annotations on mAP performance.

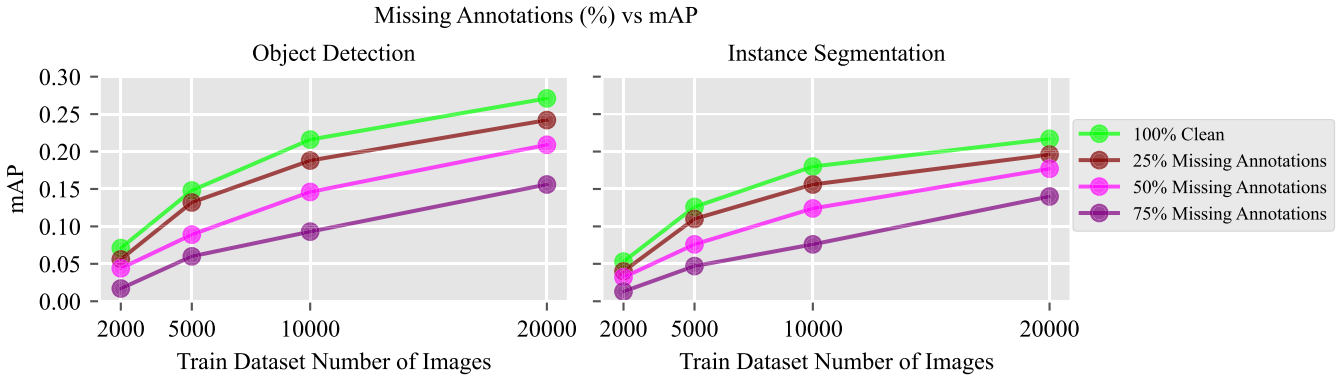

FIGURE 8. Mask-RCNN model results on ADE20K subsets for varying missing annotations on mAP performance.

Firstly, a fixed degradation amount is used across all three annotation uncertainties, with 25%, 50%, and 75% being the chosen degradation amounts. Missing annotations are first introduced into the dataset, following this, class labels are changed and finally, localization uncertainty is added to the given degradation amount. Only Mask-RCNN and YOLACT models were used for this experiment to reduce the computational workload on the COCO dataset, with only Mask-RCNN used for the ADE20K dataset. The results of the combined annotation uncertainties on the COCO dataset are seen in Fig. 12. The legend for the respective figures can be interpreted as follows: 100% Clean refers to the original ground truth annotations being used, 25% combined refers to 25% of the annotations being dropped from the dataset followed by 25% of the dataset being introduced with incorrect class labels and finally 25% of the annotations with induced localization uncertainty. Finally, the results of the ADE20K datasets can be seen in Fig. 13. The findings from this experiment highlight the importance of ground truth annotations, as when annotation uncertainties are combined the degree of degradation recorded is severe.

VI. DISCUSSION

The results enable us to investigate the relationship between annotation quality and quantity on mAP performance for object detection and instance segmentation for varying annotation qualities. As the training size increases for COCO and ADE20K's original ground truth annotations so does mAP performance as seen in Table 1. However, collecting

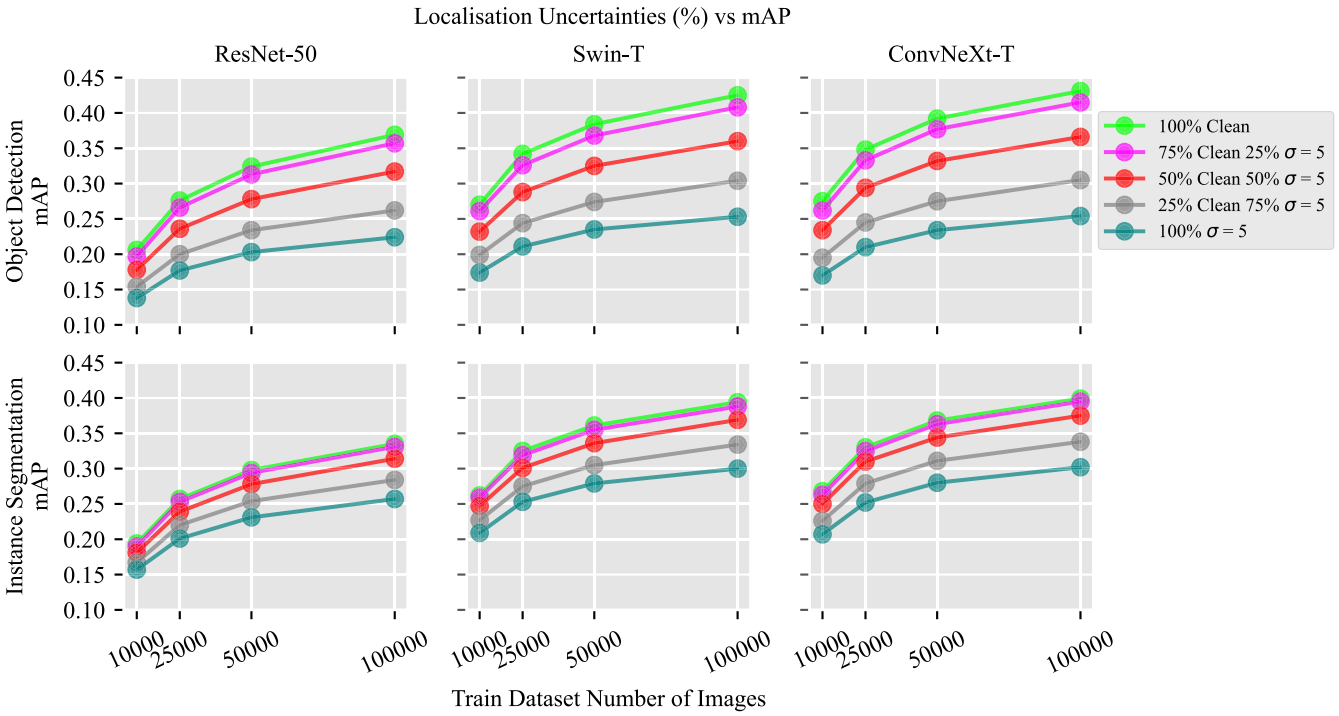

FIGURE 9. Per backbone Mask-RCNN COCO results for varying localization uncertainty on mAP performance.

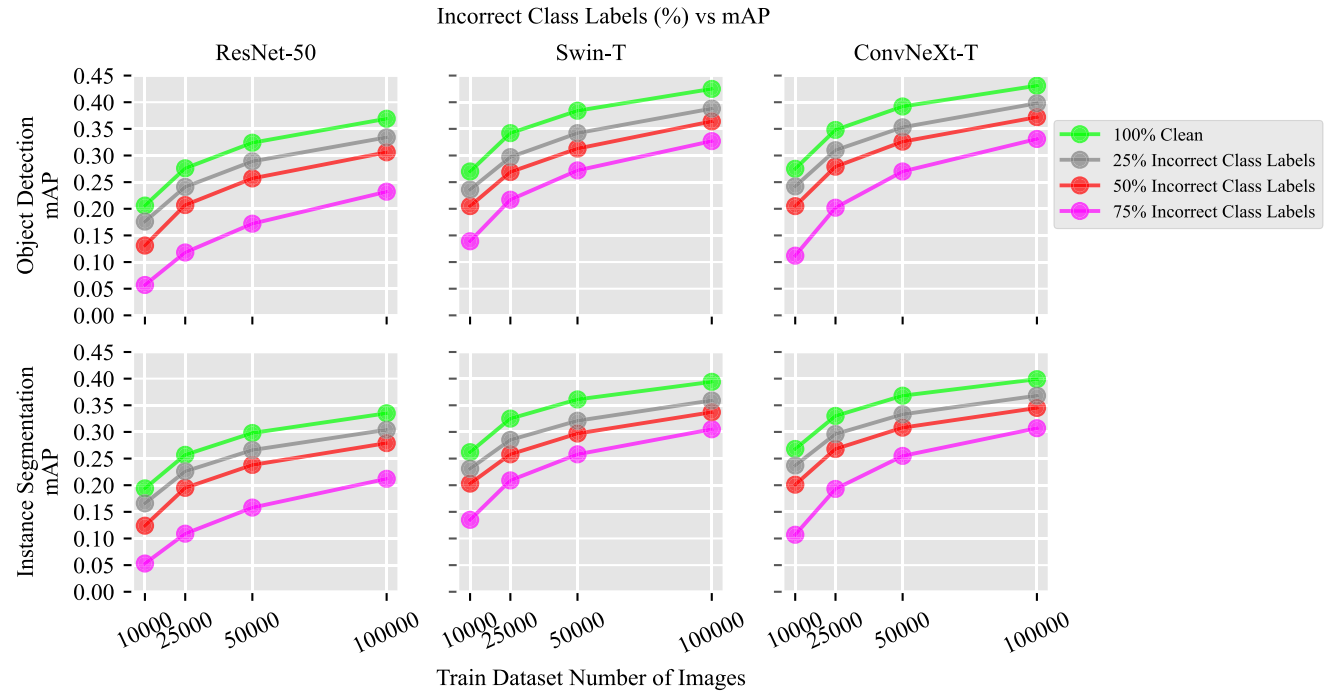

FIGURE 10. Per backbone Mask-RCNN COCO results for varying missing annotations on mAP performance.

and annotating object detection and instance segmentation datasets is a labor-intensive task that carries both a time and cost investment. In practice, due to the large dataset sizes needed for fully-supervised computer vision tasks, this commonly leads to suboptimal quality of bounding box and polygon mask annotations.

A. LOCALIZATION UNCERTAINTY

Results of the localization uncertainty experiment as shown in Fig. 3 & 4, suggest the utility of additional data is directly related to its annotation quality. Increasing the percentage of the annotations with degradation introduced results in a drop in mAP performance. This is also true for the

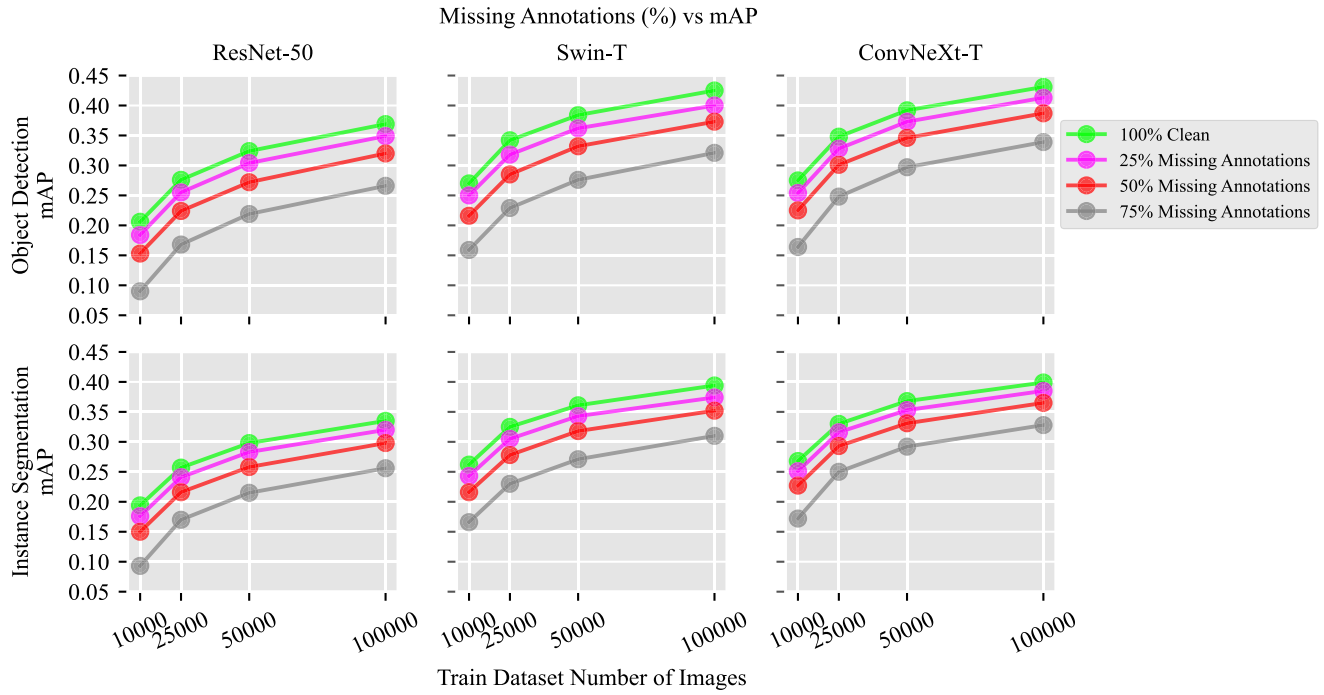

FIGURE 11. Per backbone Mask-RCNN COCO results for varying missing annotations on mAP performance.

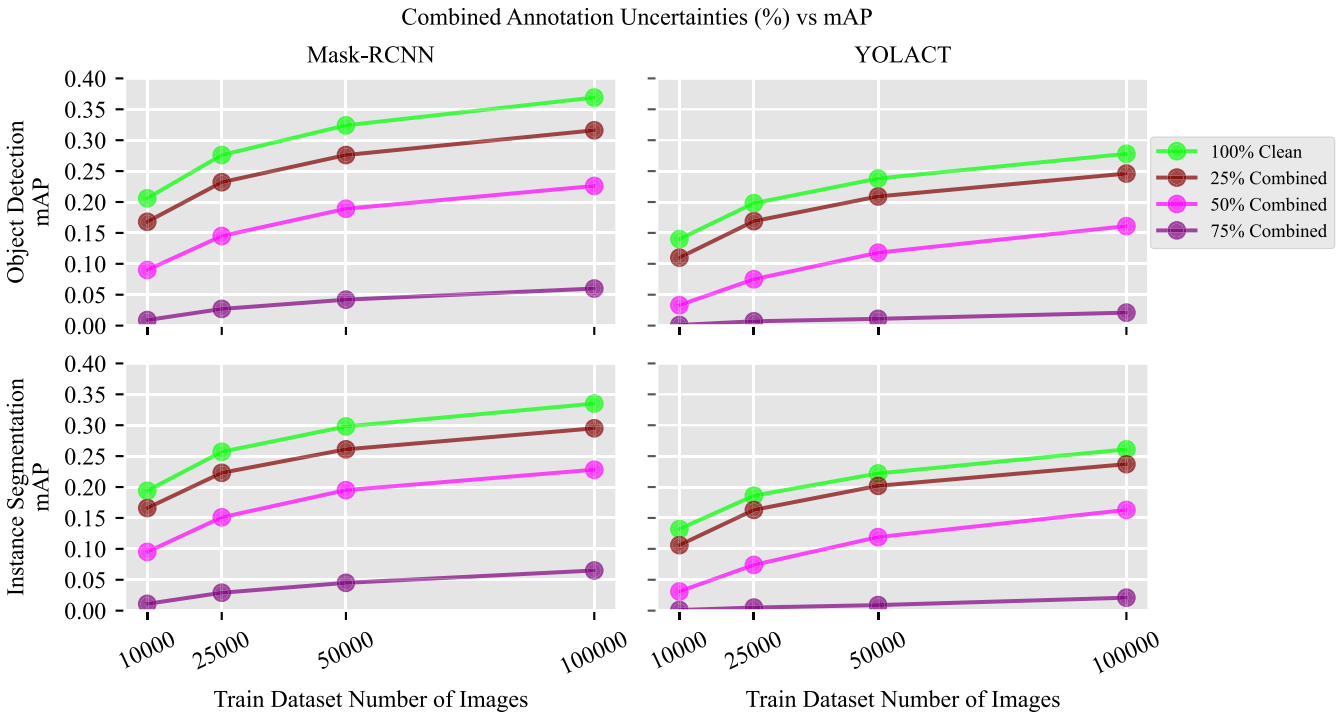

FIGURE 12. COCO results for varying combined annotation uncertainties on mAP performance.

level of uncertainty introduced. As the standard deviation of the introduced noise (σ) increases, the annotations experience greater degradation, leading to a more significant decrease in mAP. For reference, increasing the training dataset size by a factor of 10 when using the original ground truth COCO annotations, from 10,000 training images to 100,000, has resulted in an increase in mAP performance of 0.142 and 0.129 for object detection and instance segmentation, respectively, when averaging across all models. However, the quality of the data is important.

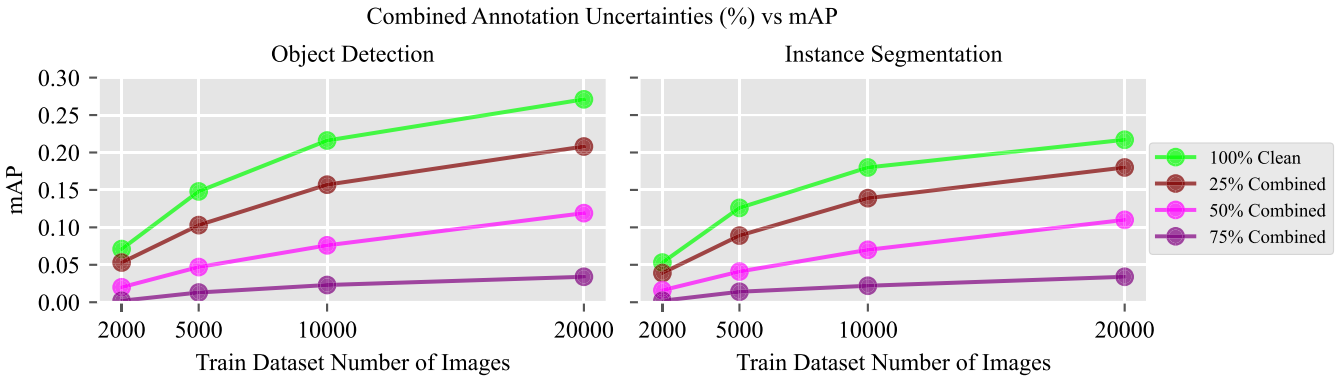

FIGURE 13. Mask-RCNN model results on ADE20K subsets for varying combined annotation uncertainties on mAP performance.

TABLE 1. mAP performance per training sample size.

Task	Baseline Training Images	2.5x Training Images	5x Training Images	10x Training Images
COCO				
Mask-RCNN				
Object Detection	0.206	0.276	0.324	0.369
Instance Segmentation	0.194	0.257	0.298	0.335
YOLACT				
Object Detection	0.140	0.198	0.238	0.278
Instance Segmentation	0.132	0.186	0.222	0.261
Mask2Former				
Object Detection	0.280	0.346	0.376	0.406
Instance Segmentation	0.276	0.336	0.362	0.392
ADE20K				
Mask-RCNN				
Object Detection	0.206	0.276	0.324	0.369
Instance Segmentation	0.194	0.257	0.298	0.335

TABLE 2. Comparison of τ across all 3 annotation uncertainty types on the COCO dataset.

Annotation Uncertainty Type	25% Uncertainty Rate	50% Uncertainty Rate	75% Uncertainty Rate
Object Detection			
Localization Error $\sigma = 2$	99.9	98.7	95.8
Localization Error $\sigma = 5$	97.9	91.2	82.3
Incorrect Class Labels	**89.2**	**75.1**	**48.9**
Missing Annotations	92.9	82.4	62.9
Instance Segmentation			
Localization Error $\sigma = 2$	99.9	99.2	97.2
Localization Error $\sigma = 5$	98.7	93.7	86.4
Incorrect Class Labels	**88.4**	**74.4**	**47.9**
Missing Annotations	93.0	83.2	64.3

Training on a dataset of 100,000 images with the maximum degradation used for all annotations that is $\sigma = 5$ for 100% of the annotations, results in an average gain in mAP performance across the 3 models used of 0.044 for object detection and 0.049 for instance segmentation, relative to 10,000 images with the original ground truth annotations. The results for varying the backbones can be seen in Fig. 9, with the trends being consistent across all 3 backbones used. The relationship between mAP degradation when considering sample size, τ, can be seen in Fig. 14. The results allow for a more direct comparison of how degradation amounts affect mAP performance per sample size. While the degradation amounts are introduced as percentages of the sample size, increasing the sample size does not drown out the effect of degradation annotations for localization uncertainty. This is reflected by the degradation amount remaining approximately constant throughout each sample size for both object detection and instance segmentation. The results from the localization uncertainty experiment suggest that perfectly labeled bounding boxes and polygon masks achieve the best scores per sample size. However, some small levels of localization uncertainty are not detrimental to model

TABLE 3. Breakdown of number of annotations for missing annotations datasets.

Percentage of Missing Annotations	Baseline Training Images No. of Annotations	2.5x Training Images No. of Annotations	5x Training Images No. of Annotations	10x Training Images No. of Annotations
COCO				
0	71,948	180,697	359,714	719,325
25	53,819	135,681	269,933	539,448
50	35,960	90,068	179,431	359,654
75	17,869	45,322	89,933	180,323
ADE20K				
0	19,275	48,526	96,535	192,788
25	14,459	36,521	72,565	144,222
50	9,552	24,166	48,292	96,405
75	4,778	12,065	24,149	48,366

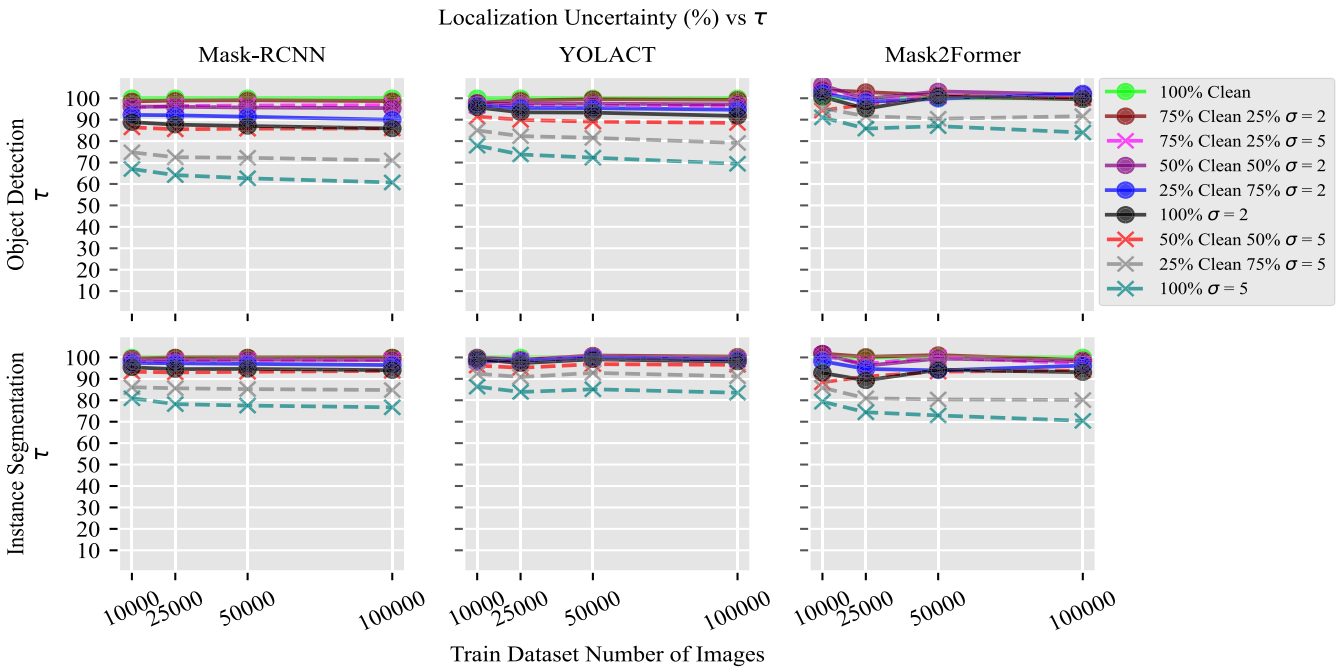

FIGURE 14. COCO results for varying localization uncertainty on τ.

performance. This is reflected by an average decrease in mAP performance across all models and the four subsets of 0.003 and 0.0003 for 75% of the original dataset untouched and 25% with $\sigma = 2$ introduced into the annotations for object detection and instance segmentation, respectively.

B. INCORRECT CLASS LABELS

Results of the effect of incorrect class labels as shown in Fig. 5 & 6 show that, as the number of incorrect class labels increases, the degradation in the models' mAP performance becomes more pronounced. Training on the dataset of 100,000 images with the three incorrect class label percentages of 75%, 50%, and 25% results in a decrease of mAP performance of 0.12, 0.05, and 0.024 respectively, for object detection when averaged across all 3 models on the COCO dataset. For instance segmentation the decreases in mAP performance are 0.117, 0.048, and 0.026. These results suggest that incorrect class labels degrade model performance approximately the same for object detection and instance segmentation. The degradation in mAP performance is also not linear as the difference in mAP performance from 75% to 50% is 0.069 while the difference from 50% to 25% is 0.024 when averaging the differences between object detection and instance segmentation. The results for varying the backbones can be seen in Fig. 15, with the trends being consistent across all 3 backbones used. The relationship between mAP degradation when considering sample size, τ, can be seen in Fig. 15. For incorrect class labels, it appears increasing the sample size has the ability to reduce the effect of degraded annotations for 50% and 75% of incorrect class labels. This is reflected by an upward trend in τ as sample size increases in Fig. 15. However, the effect is not as strong for 25% of incorrect class labels. The results from the incorrect class labels experiment suggest due diligence is required when annotating class labels as this can significantly degrade mAP performance.

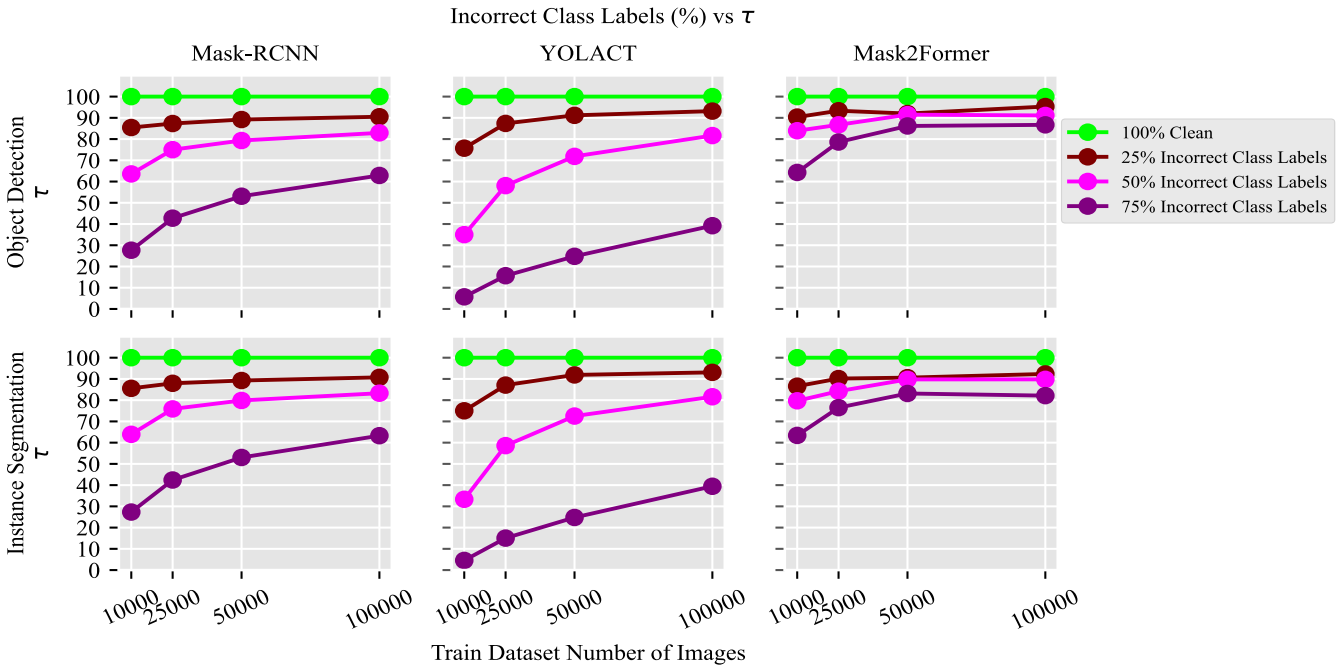

FIGURE 15. COCO results for varying incorrect class labels on τ.

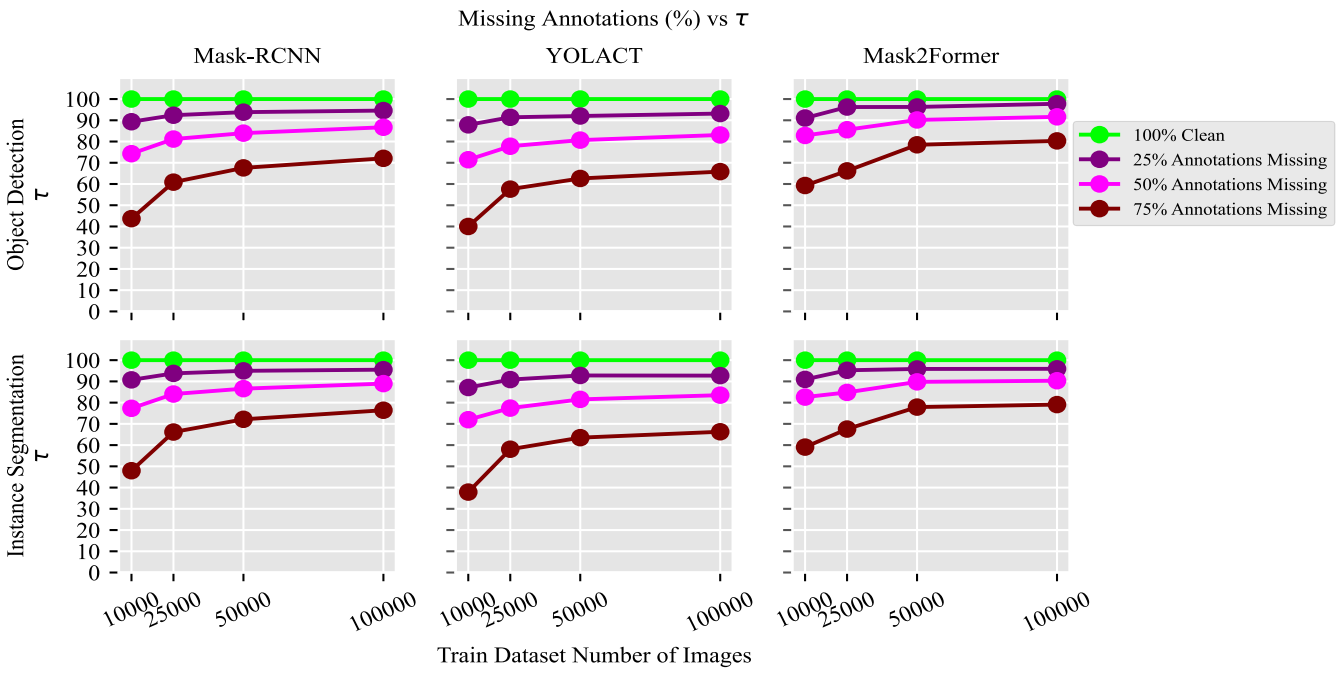

FIGURE 16. COCO results for varying missing annotations on τ.

C. MISSING ANNOTATIONS

Results of the effect of missing annotations, as shown in Fig. 7 & 8, show that, as we remove more of the annotations from the dataset, this results in a decrease in mAP performance. Training on the dataset of 100,000 images with the three missing annotation percentages of 75%, 50%, and 25% results in a decrease in mAP performance of 0.093, 0.043, and 0.016 respectively, for object detection when averaged across all 3 models on the COCO dataset. For instance segmentation, the differences in mAP performance are 0.083, 0.039, and 0.017. These results suggest that missing annotations degrade model performance marginally more for object detection than

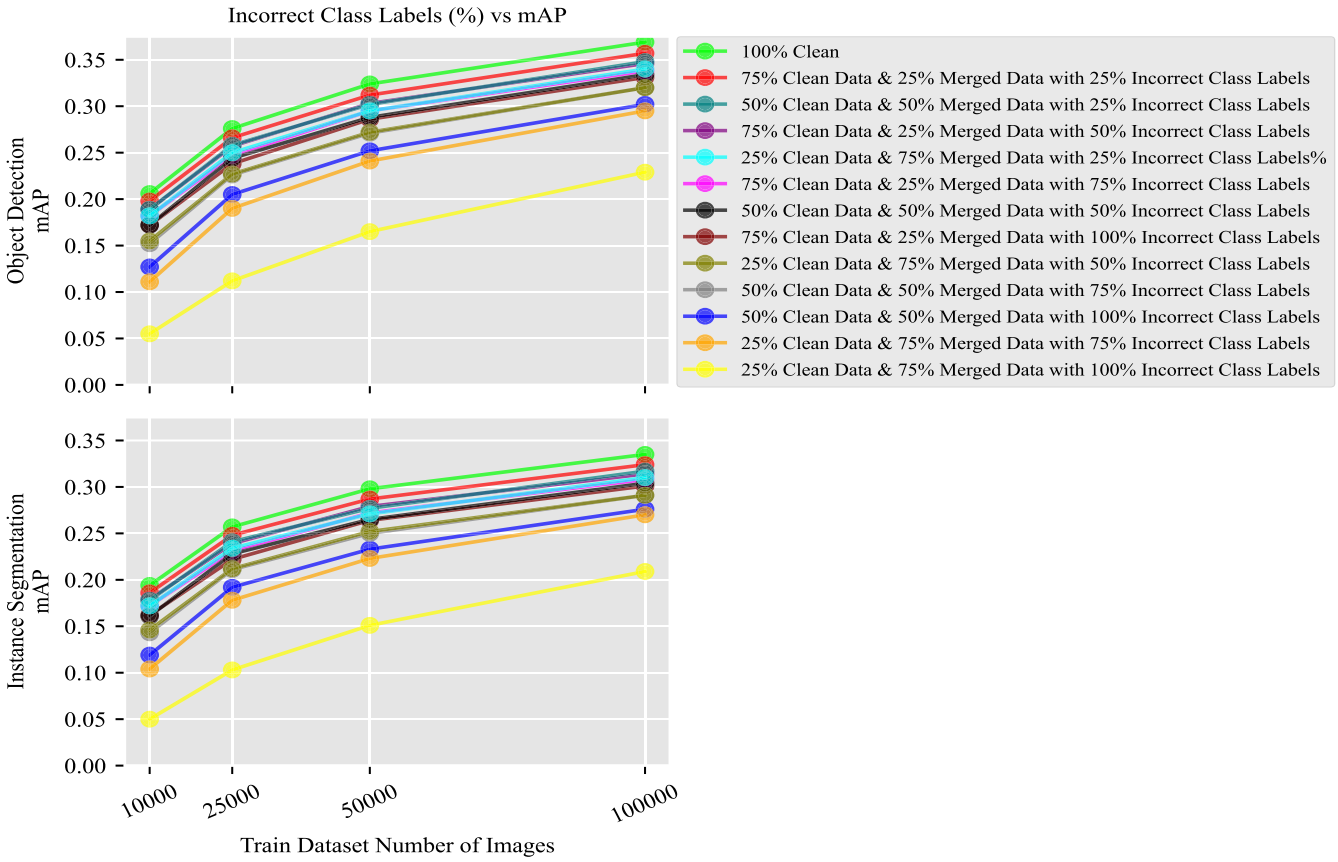

FIGURE 17. COCO mask-RCNN results for varying incorrect class labels when merging.

instance segmentation. The degradation in mAP performance is also not linear as the difference in mAP performance from 25% to 50% is 0.026, whilst the difference from 50% to 75% is 0.048 when averaging the differences between object detection and instance segmentation. The results for varying the backbones can be seen in Fig. 11, with the trends consistent across all 3 backbones used. The relationship between mAP degradation when considering sample size, τ, can be seen in Fig. 16. For missing annotations, it appears increasing the sample size has the ability to reduce the effect of degraded annotations for 75% and 50% of missing annotations. This is reflected by an upward trend in τ as sample size increases in Fig. 16. The results from the effect of missing annotations suggest the number of annotations is more critical than the number of images in your dataset. This reinforces the need to utilize your dataset fully, annotating each object of interest before collecting more data.

D. COMBINED ANNOTATION UNCERTAINTIES

The results of combining the annotation uncertainties can be seen in Fig. 12 & 13. The results show that as we compound the annotation uncertainties, this further degrades mAP performance in comparison to their individual parts, with 75% of combined annotation uncertainties for each sample size scoring an approximate mAP < 0.07 for both models used. The results highlight the importance of annotation quality given the degradation amounts when combining the uncertainties.

E. INTRODUCTION OF ERRORS

The results of comparing a percentage of annotations having annotation uncertainty introduced in comparison to merging a clean and second dataset of lesser quality can be seen in Fig. 19 & 20. The results show that for both object detection and instance segmentation, the difference in mAP performance between these two methods is negligible when considering localization uncertainty and incorrect class labels. For the models, it appears degraded annotations have the same effect on mAP performance regardless of how they were introduced into the dataset. To illustrate, consider a scenario with 100 annotations distributed over 10 images, with each image containing 10 annotations. The situation in which each of the 10 images has one degraded annotation is equivalent to having 9 images with correct annotations and one image with all annotations incorrect. As merging a dataset with no annotations adds no value to the dataset, there is no direct comparison between a percentage of annotations having missing annotations and merging a clean and second dataset of lesser quality.

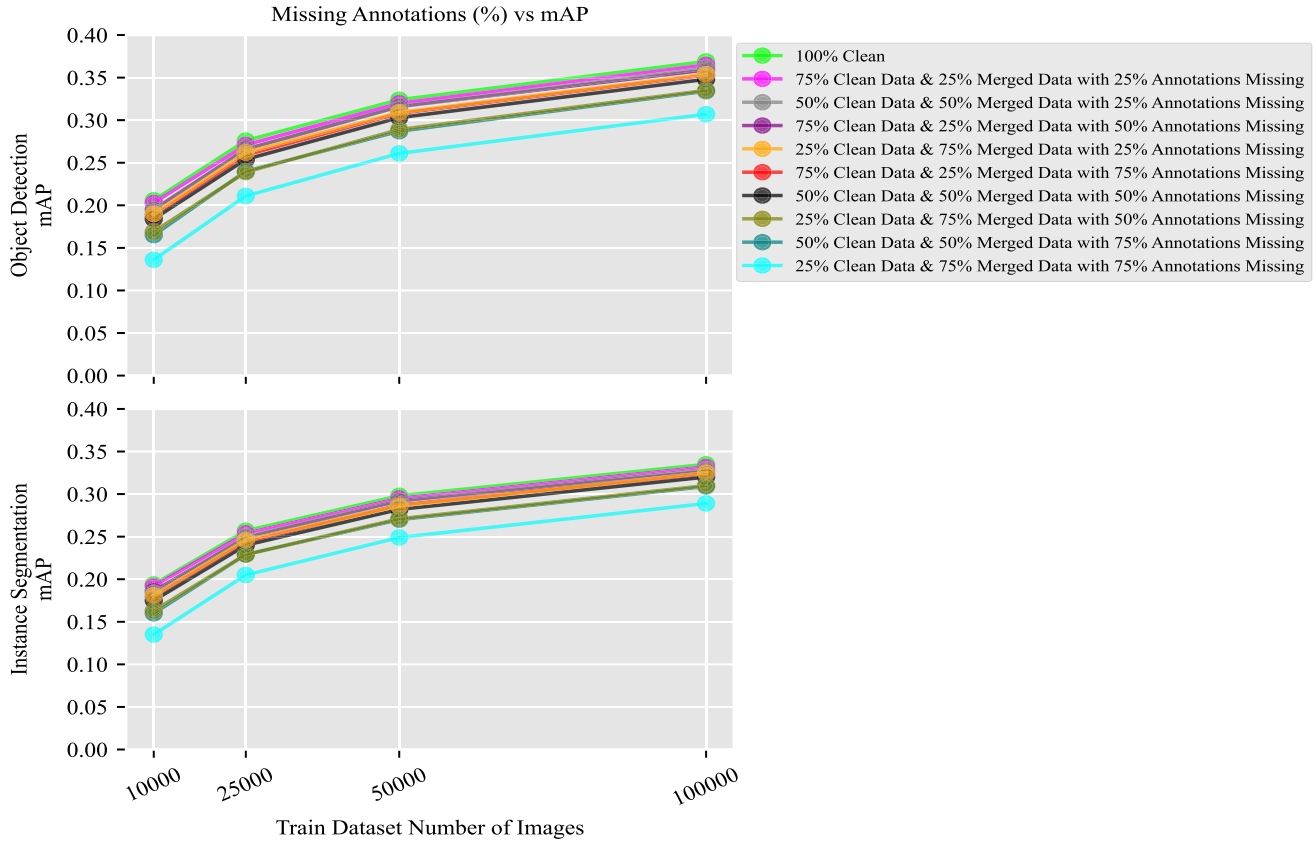

FIGURE 18. COCO mask-RCNN results for varying missing annotations when merging.

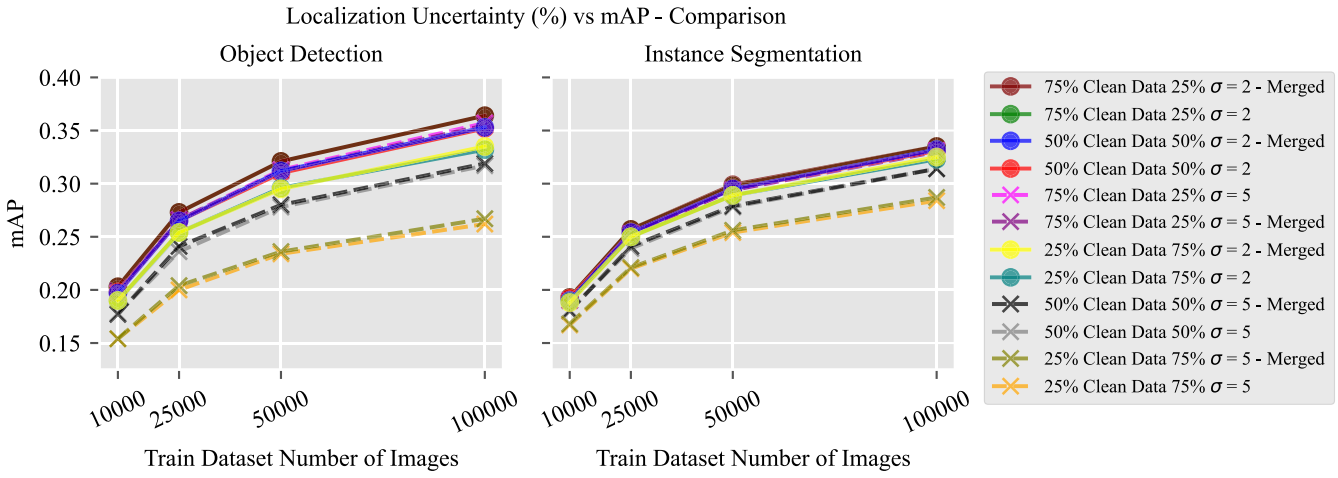

FIGURE 19. Mask-RCNN - Comparison of thresholding & merging a second dataset with localization uncertainty on mAP performance.

F. DEGRADATION OF mAP PERFORMANCE

As expected, for all 3 types of annotation uncertainty introduced into the datasets, there is a reduction in mAP performance as shown in Table 2. However, each type of annotation uncertainty has varying impacts on the magnitude of the degradation of mAP performance. Table 2 suggests incorrect class labels have the highest degradation effect on mAP performance followed by missing annotations and lastly localization errors. Whilst annotation errors have a negative effect on mAP performance, the results from each of the experiments show that gains can be made from the use of additional data of lesser annotation quality. One potential reason as to why localization errors retain high τ values is due to the metric chosen. As mAP is calculated over a range

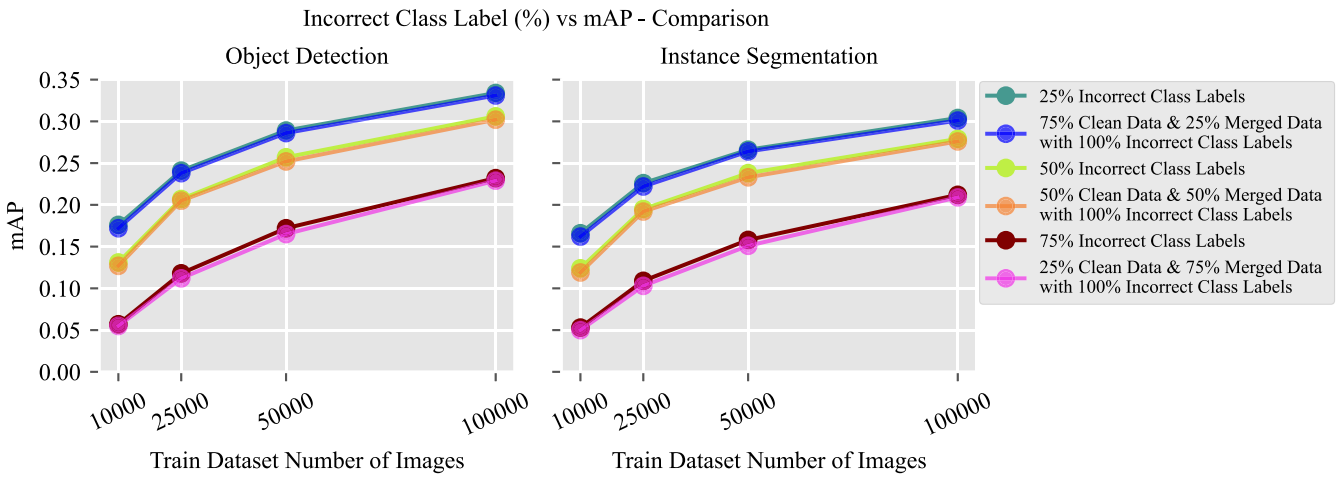

FIGURE 20. Mask-RCNN - Comparison between thresholding and merging a second dataset with incorrect class labels on mAP performance.

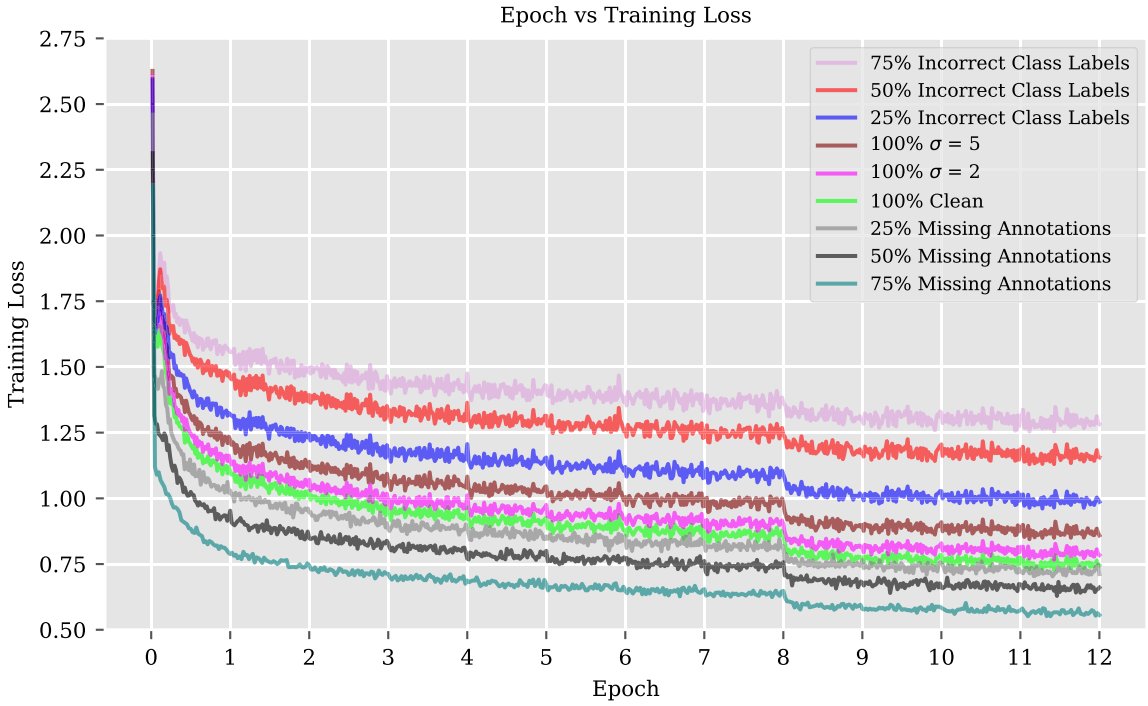

FIGURE 21. Learning curves for various mask-RCNN models trained on 100,000 images.

of intersection over union values ranging from 0.5 to 0.95, small localization uncertainties may not be detrimental to the overall performance. Furthermore, these results are in line with our previous paper [10], which showed when the entire dataset has localization uncertainty introduced, we observe a degradation in mAP score of 0.0241 & 0.0135 for each unit increase in σ for object detection and instance segmentation respectively. For this work, since the localization uncertainty is introduced as a percentage of the dataset, we expect the degradation amount to decrease.

The learning curves for the various Mask-RCNN models trained on 100k images can be seen in Fig. 21. The training losses are higher for models trained with incorrect class labels and localization uncertainties in comparison to the model trained with the original ground truth annotations. In contrast, the models with missing annotations can achieve lower training errors. This is possible as models can better fit the smaller training datasets. However, all models follow the same trends for convergence.

G. QUALITATIVE ANALYSIS

From inspection of the inference results as seen in Fig. 22, missing annotations and incorrect class labels both reduce the

FIGURE 22. Inferences of various Mask2Former models trained on 100,000 images.

confidence scores of predictions in comparison to the baseline model, which was trained with the original ground truth annotations. The results suggest that missing annotations and incorrect class labels in your dataset hinder the model's ability to converge on the optimal weights. This, in turn, reduces the confidence scores of each inference, resulting in predictions being filtered out due to confidence scores being lower than the allowable threshold. One potential reason for incorrect class labels degrading mAP performance the most is due to the confusion they introduce, an outcome that is supported by [44]. As the model is trained, the same class object is shown with multiple class labels throughout the dataset. This prevents the model from converging on the understanding of the objects of interest. As for missing annotations, the induced missing annotations in the dataset only affect the class objectness score, resulting in objects of interest being considered the background class during training [45]. This may bias the model into predicting the class only in certain circumstances; however, there is no confusion with other classes. Concerning data collection and

annotation, the results suggest that for optimal performance for both object detection and instance segmentation, the dataset should be perfectly labeled. However, considering the investment needed for annotating large-scale datasets, a trade-off between quality and quantity may be of interest. Annotating a larger dataset with small levels of annotation uncertainty may yield better performance than a smaller dataset that is perfectly labeled.

H. LIMITATIONS & FUTURE WORK

The findings of this study have to be seen in light of some limitations. The use of a discrete space of 10,000, 25,000, 50,000, and 100,000 training images for the COCO dataset and 2,000, 5,000, 10,000, and 20,000 for the ADE20K dataset along with the degradation percentages of 25/50/75/100 were used to limit the number of models that had to be trained and thus limit compute resources. A finer scale for both dataset size and degradation amounts could provide more detailed insight as to when the data of lesser annotation quality begin to negatively affect mAP performance. Experiments on further object detection and instance segmentation datasets along with different model architectures should be conducted in the future to further develop and validate these initial findings and determine whether the results from these experiments generalize. In addition to this, an investigation into the relationship between annotation quality and quantity for further computer vision tasks, such as image classification or semantic segmentation, would be of interest to further develop the understanding of this relationship. Furthermore, it is important to acknowledge that the methods used to introduce annotation uncertainties may in themselves carry biases. Lastly, it would be of interest to investigate the effects of data augmentation and its ability to add diversity to the dataset in comparison to adding new data.

VII. CONCLUSION

In this paper, the relationship between annotation quality and quantity and their effects on mAP performance for both object detection and instance segmentation is studied. Datasets of training image sizes of 10,000, 25,000, 50,000, and 100,000 were created from the COCO 2017 detection dataset for these experiments, in addition to 2,000, 5,000, 10,000, and 20,000 from the ADE20K dataset. To investigate the effect of suboptimal annotations along with the relationship between annotation quality and quantity, three different types of annotation uncertainties were introduced to the ground truth annotations of the datasets: localization uncertainty, incorrect class labels, and missing annotations. Mask-RCNN, YOLACT, and Mask2Former models were trained on the induced error datasets. In total, 456 models were trained to investigate the relationship between annotation quality and quantity for varying levels of annotation uncertainty over the four dataset sizes.

The results show that all three annotation uncertainties negatively affect mAP performance. The degree to which each type of annotation uncertainty degrades mAP differs as seen in Table 2, with incorrect class labels degrading the mAP performance the most, followed by missing annotations and lastly, localization uncertainty. While the results show that perfectly labeled data outperforms degraded annotations for a fixed sample size, there is utility in adding additional data of lesser annotation quality. The extent of the benefits of the additional data is directly related to how degraded the annotations' are. The results also suggest that degraded annotations have the same impact on mAP performance irrespective of how the annotations were introduced into the dataset.

This study has investigated the relationship between annotation quality and quantity for mAP performance for the tasks of object detection and instance segmentation, on four subsets of the COCO and ADE20K datasets. The reduction in mAP performance for all three annotation uncertainties reinforces the importance of accurate annotations for both fully supervised object detection and instance segmentation tasks. Furthermore, the results inform us of how each type of annotation uncertainty impacts mAP performance in addition to providing insight into the relationship between annotation quality and quantity on mAP performance for object detection and instance segmentation.

REFERENCES

[1] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler, "DatasetGAN: Efficient labeled data factory with minimal human effort," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10145–10155.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 740–755.

[4] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8429–8438.

[5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023, *arXiv:2304.02643*.

[6] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 843–852.

[7] B. Sabiri., B. El Asri., and M. Rhanoui., "Mechanism of overfitting avoidance techniques for training deep neural networks," in *Proc. 24th Int. Conf. Enterprise Inf. Syst. (ICEIS)*, vol. 1. Setúbal, Portugal: SciTePress, 2022, pp. 418–427.

[8] Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6725–6735.

[9] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14408–14419.

[10] C. Agnew, C. Eising, P. Denny, A. Scanlan, P. Van De Ven, and E. M. Grua, "Quantifying the effects of ground truth annotation quality on object detection and instance segmentation performance," *IEEE Access*, vol. 11, pp. 25174–25188, 2023.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[12] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Proc. Workshops 26th AAAI Conf. Artif. Intell.*, 2012, pp. 1–7.
[13] COCO. (2021). *Detection Evaluation Metrics*. Accessed: Oct. 18, 2022. [Online]. Available: https://cocodataset.org/#detection-eval
[14] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.
[15] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.
[16] N. Jaipuria, X. Zhang, R. Bhasin, M. Arafa, P. Chakravarty, S. Shrivastava, S. Manglani, and V. N. Murali, "Deflating dataset bias using synthetic data augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 772–773.
[17] C. Liang, Z. Yang, L. Zhu, and Y. Yang, "Co-learning meets stitch-up for noisy multi-label visual recognition," *IEEE Trans. Image Process.*, vol. 32, pp. 2508–2519, 2023.
[18] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
[19] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2304–2313.
[20] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4334–4343.
[21] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[22] Y. Xu, L. Zhu, Y. Yang, and F. Wu, "Training robust object detectors from noisy category labels and imprecise bounding boxes," *IEEE Trans. Image Process.*, vol. 30, pp. 5782–5792, 2021.
[23] H. Li, Z. Wu, C. Zhu, C. Xiong, R. Socher, and L. S. Davis, "Learning from noisy anchors for one-stage object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10588–10597.
[24] J. Mao, Q. Yu, and K. Aizawa, "Noisy localization annotation refinement for object detection," *IEICE Trans. Inf. Syst.*, vol. 104, no. 9, pp. 1478–1485, 2021.
[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learn. Represent. Learn. Workshop*, 2015, pp. 1–9.
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[28] S. Shahinfar, P. Meek, and G. Falzon, "'How many images do I need?' Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring," *Ecol. Inform.*, vol. 57, May 2020, Art. no. 101085.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[32] J. Ma, Y. Ushiku, and M. Sagara, "The effect of improving annotation quality on object detection datasets: A preliminary study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4850–4859.
[33] M. Xu, Y. Bai, and B. Ghanem, "Missing labels in object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 1–10.
[34] L. Yang, F. Meng, H. Li, Q. Wu, and Q. Cheng, "Learning with noisy class labels for instance segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 38–53.
[35] V. Taran, Y. Gordienko, A. Rokovyi, O. Alienin, and S. Stirenko, "Impact of ground truth annotation quality on performance of semantic image segmentation of traffic conditions," in *Advances in Computer Science for Engineering and Education II*. Cham, Switzerland: Springer, 2020, pp. 183–193.
[36] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019.
[37] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
[38] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9157–9166.
[39] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1290–1299.
[40] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[43] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.
[44] A. Ghiassi, R. Birke, and L. Y. Chen, "Multi label loss correction against missing and corrupted labels," in *Proc. Asian Conf. Mach. Learn.*, 2023, pp. 359–374.
[45] H. Zhang, F. Chen, Z. Shen, Q. Hao, C. Zhu, and M. Savvides, "Solving missing-annotation object detection with background recalibration loss," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1888–1892.

CATHAOIR AGNEW received the B.S. degree in financial mathematics and the M.S. degree in artificial intelligence and machine learning from the University of Limerick, Limerick, Ireland, in 2020 and 2021, respectively, where he is currently pursuing the Ph.D. degree in electronic and computer engineering. His research interests include artificial intelligence and computer vision.

ANTHONY SCANLAN received the B.Sc. degree in experimental physics from the National University of Ireland, Galway, Ireland, in 1998, and the M.Eng. and Ph.D. degrees in electronic engineering from the University of Limerick, Limerick, Ireland, in 2001 and 2005, respectively. He is currently a Senior Research Fellow with the Department of Electronic and Computer Engineering, University of Limerick, and has been a Principal Investigator for several research projects in the areas of signal processing and data converter design. His current research interests include artificial intelligence, computer vision, and their industrial and environmental applications.

PATRICK DENNY (Member, IEEE) received the B.Sc. degree in experimental physics and mathematics from NUI Maynooth, Ireland, in 1993, and the M.Sc. degree in mathematics and the Ph.D. degree in physics from the University of Galway, Ireland, in 1994 and 2000, respectively. He was with GFZ Potsdam, Germany. From 1999 to 2001, he was a RF Engineer with AVM GmbH, Germany, developing the RF hardware for the first integrated GSM/ISDN/USB modem. After working in super-computing with Compaq-HP, from 2001 to 2002, he joined Connaught Electronics Ltd. (later Valeo), Galway, as the Team Leader of RF Design. Over the next 20 years, he was the Lead Engineer developing novel RF and imaging systems and led the development of the first mass-production HDR automotive cameras for leading car companies, including Jaguar Land Rover, BMW, and Daimler. In 2010, he became an Adjunct Professor of engineering and informatics with the University of Galway. In 2022, he became a Lecturer in artificial intelligence with the Department of Electronic and Computer Engineering, University of Limerick, Ireland, where he became an Associate Professor of artificial intelligence and imaging with the Department of Computer Science and Information Systems (CSIS), in 2024. He is currently the Co-Founder and a Committee Member of the IEEE P2020 Automotive Imaging Standards Group, the AutoSens Conference on Automotive Imaging, and the IS&T Electronic Imaging Autonomous Vehicles and Machines (AVM) Conference.

EOIN M. GRUA was born in Cork, Ireland, in 1993. He received the B.S. degree in liberal arts and sciences from Amsterdam University College, Amsterdam, The Netherlands, in 2015, the M.S. degree in computer science from Swansea University, Swansea, Wales, in 2016, and the Ph.D. degree in computer science from Vrije Universiteit Amsterdam, Amsterdam, in 2021. In 2021, he was a Research Assistant with the University of Limerick, Limerick, Ireland. He is currently an Assistant Professor with the Department of Electronic and Computer Engineering, University of Limerick. His research interests include artificial intelligence, software engineering and architecture, and sustainability.

PEPIJN VAN DE VEN received the M.Sc. degree in electronic engineering from the Eindhoven University of Technology, The Netherlands, in 2000, and the Ph.D. degree in artificial intelligence for autonomous underwater vehicles from the University of Limerick, in 2005. He is currently a Professor in artificial intelligence with the University of Limerick. His research interests include artificial intelligence and machine learning, with a particular focus on medical applications.

CIARÁN EISING (Senior Member, IEEE) received the B.E. degree in electronic and computer engineering and the Ph.D. degree from the National University of Ireland, Galway, in 2003 and 2010, respectively. From 2009 to 2020, he was the Computer Vision Team Lead and an Architect with Valeo Vision Systems, where he also held the title of a Senior Expert. In 2016, he was awarded the position of an Adjunct Lecturer with the National University of Ireland. In 2020, he joined the University of Limerick as an Artificial Intelligence and Computer Vision Lecturer.