Marcel P. Schilling*, Niket Ahuja, Luca Rettenberger, Tim Scherr, and Markus Reischl

Impact of Annotation Noise on Histopathology Nucleus Segmentation

https://doi.org/10.1515/cdbme-2022-1051

Abstract: Deep learning is often used for automated diagnosis support in biomedical image processing scenarios. Annotated datasets are essential for the supervised training of deep neural networks. The problem of consistent and noise-free annotation remains for experts such as pathologists. The variability within an annotator (intra) and the variability between annotators (inter) are current challenges. In clinical practice or biology, instance segmentation is a common task, but a comprehensive and quantitative study regarding the impact of noisy annotations lacks. In this paper, we present a concept to categorize and simulate various types of annotation noise as well as an evaluation of the impact on deep learning pipelines. Thereby, we use the multi-organ histology image dataset MoNuSeg to discuss the influence of annotator variability. We provide annotation recommendations for clinicians to achieve high-quality automated diagnostic algorithms.

Keywords: Instance Segmentation, Annotator Variability, Deep Learning, Image Processing

1 Introduction

Deep Neural Networks (DNNs) achieve high-quality results in complex image processing tasks, i.e., cell segmentation in biology or medicine [1, 4]. Though, supervised DNNs require annotated images by experts such as pathologists or biologists.

Noisy annotations, i.e., partially false or wrong annotations are an open problem. In general, annotation noise can be divided into variability within an annotator (intra) and between different annotators (inter) [3]. There are various reasons for annotation noise like different levels of knowledge or experience of annotators, fatigue, limited time budget, varying decision boundaries, or different hardware devices used during annotation. In clinical practice, division of labor is common to reduce the effort for individual annotators. However, annotation noise in terms of consistency issues may affect the dataset.

Karimi et al. [3] give a general overview w.r.t. noisy annotation in image processing. Various methods to handle noisy annotations, i.e., detecting and correcting or reducing the impact during training of DNNs, are discussed. In addition, the impact of noisy annotation for *classification* [6] or *semantic segmentation* [8, 9] is investigated. Löffler et al. [5] examine the impact of wrong segmentation results w.r.t. the performance of tracking algorithms. Following the arguments in [3], biomedical problems, often consisting of only a small number of samples, are prone to noisy annotations. Northcutt et al. [6] argue that noisy annotations are disadvantageous since there could be negative effects on the DNN training or metrics may become unreliable yielding perhaps a wrong model selection.

However, less consideration is given to *instance segmentation*, although, this problem often occurs in biomedical DL applications. A concept for a comprehensive study w.r.t. annotation noise in the case of instance segmentation is missing. Recommendations for experts regarding annotation noise, what to pay special attention to during annotation to get a high quality DNN, can be very useful. Further, for data scientists, the question of whether more accurate annotations can improve DNN performance remains open.

Hence, we propose a detailed categorization of noisy annotation in instance segmentation tasks utilizing a synthetic dataset. Further, we present methods for simulating intra- and inter-annotator variability. Our developed concept is completed by examining the influence of annotator variability on DL pipelines. Besides, we implement the concept and contribute an evaluation of annotator variability w.r.t. the multi-organ histology image dataset MoNuSeg [4] dealing with nucleus instance segmentation. Hence, we provide recommendations for researchers, i.e. clinicians or biologists, what is needed to pay special attention to obtain high-performing DNNs.

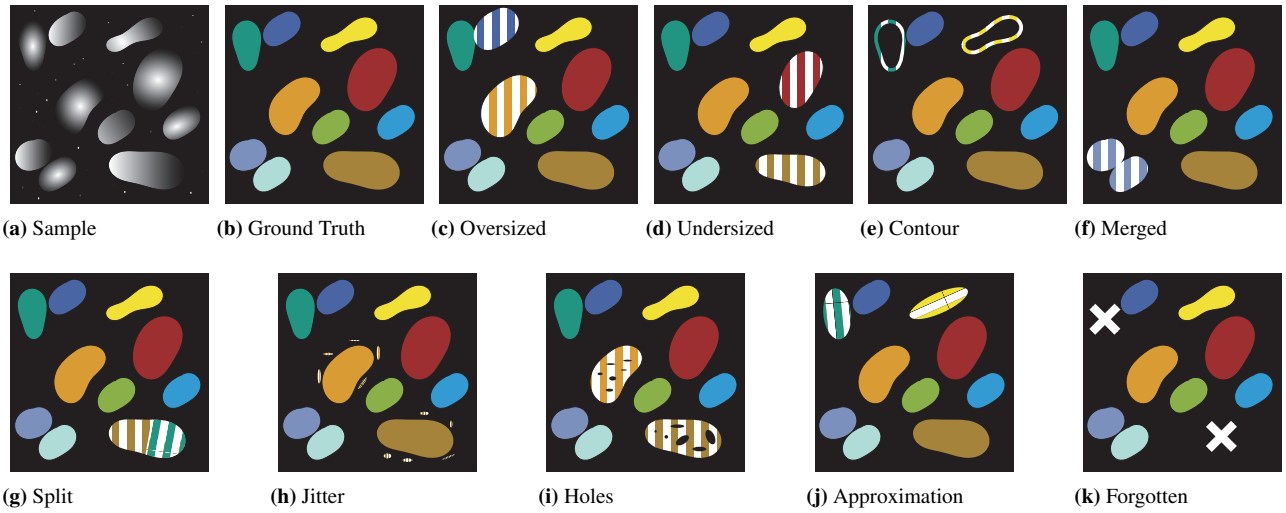
2 Material and Methods

2.1 Categorization of Annotation Noise in Instance Segmentation

First, we comprehensively categorize annotation noise in instance segmentation using a synthetic dataset in combination with various corruptions as shown in Figure 1. Our work builds on the ideas concerning annotation noise in semantic segmentation presented in [8].

Annotators may draw "oversized" or "undersized" masks. The corruption "contour" occurs in contour-based annotation

***Corresponding author: Marcel P. Schilling,** Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany, e-mail: marcel.schilling@kit.edu



(a) Sample **(b)** Ground Truth **(c)** Oversized **(d)** Undersized **(e)** Contour **(f)** Merged

(g) Split **(h)** Jitter **(i)** Holes **(j)** Approximation **(k)** Forgotten

Fig. 1: Synthetic Dataset and Corruptions. A sample (a) and corresponding ground truth annotation (b) are given. Different instances are encoded by different colors. Corruptions are depicted in (c)-(k) using white line patterns to visualize affected instances.

tools with auto-filling in the case where no loop closure is formed. "Merging" neighboring instances or "splitting" an instance into two segments are further noise types. In addition, small "jitter" instances or "holes" within an instance can be enumerated as noise in practical annotation procedures. Neglecting the exact shape of an instance, the corruption "approximation" transforms segments into ellipses. Moreover, "forgetting" to annotate instances altogether can also be referred to as annotation noise type.

In general, all those types of annotation noise could occur randomly in terms of intra-annotator variability. However, taking inter-annotator variability into account, systematic errors matter which applies only to the subset { "oversized", "undersized", "contour", "approximation" } of corruptions.

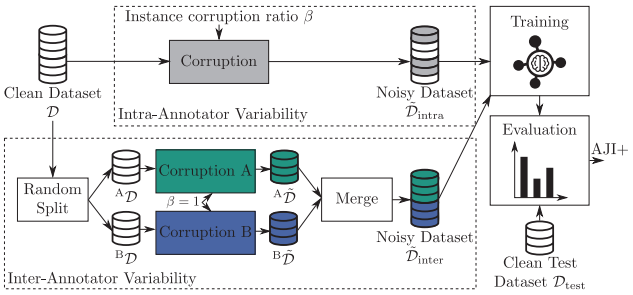


Fig. 2: Concept Overview. An initially clean dataset $\mathcal{D}$ is corrupted i) controlled by the instance corruption ratio β (intra-annotator variability) or ii) by randomly splitting, inserting two different corruption types, and merging (inter-annotator variability) leading to noisy datasets $\tilde{\mathcal{D}}_{\text{intra/inter}}$. DNNs are trained on those noisy datasets and evaluated on a clean test dataset $\mathcal{D}_{\text{test}}$ taking the performance measure AJI+ into account.

2.2 Concept for Simulating and Evaluating Annotation Noise

Figure 2 presents an overview of our proposal to investigate annotation noise in instance segmentation tasks. We assume an initially clean dataset $\mathcal{D}$ with no noise.

In the case of intra-annotator variability, corruptions are added yielding the noisy dataset $\tilde{\mathcal{D}}_{\text{intra}}$. The parameter β controls the ratio of affected instances per images, e.g. $\beta = 0.2$ in the synthetic dataset (cf. Figure 1) since two of ten instances are corrupted.

In terms of inter-annotator variability, we randomly split the dataset into two subsets $^{A}\mathcal{D}$ and $^{B}\mathcal{D}$ to simulate different annotators in the division of labor case. Different corruptions are added through which $^{A}\tilde{\mathcal{D}}$ and $^{B}\tilde{\mathcal{D}}$ result. It needs to be remarked that all instances are affected by the corruption in this scenario ($\beta = 1$). The reason for this is the consideration of systematic annotation noise. After merging the noisy dataset, $\tilde{\mathcal{D}}_{\text{inter}}$ is obtained.

Subsequently, DNNs are trained using noisy datasets. We consider a clean test dataset $\mathcal{D}_{\text{test}}$ to examine the impact of noisy annotated datasets on the DNN performance. The metric Aggregated Jaccard Index AJI+ based on the work in [2] is considered to evaluate both segmentation and detection performance of the DNN.

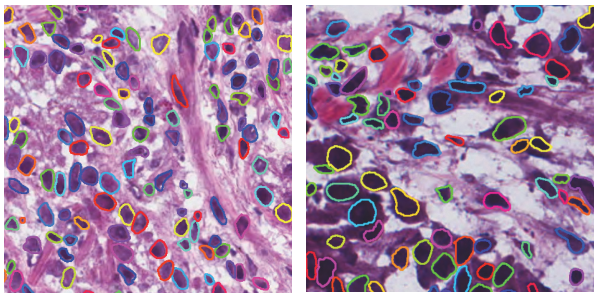


Fig. 3: MoNuSeg Dataset. Two crops including ground truth instances (contours, instances encoded by different colors) are depicted [4]. The crops originate from different organs.

3 Results

3.1 Dataset

We use our proposed concept in combination with the MoNuSeg dataset [4] to investigate the quantitative impact of annotation noise in instance segmentation tasks (cf.Figure 3). The dataset addresses the instance segmentation of cell nuclei in diverse tissue images. The training dataset is composed of 30 hematoxylin and eosin histology images (1000px x 1000px) of seven different organs including 21.623 nuclei. The test dataset covers 14 images.

3.2 Architecture, Training, and Implementation

We use a U-Net [7] to predict Euclidean distance maps with subsequent seed-based watershed post-processing for segmenting instances. Smooth L1 loss serves as an objective function for the Adam optimizer. Early stopping and learning rate scheduling are considered. The used random train-validation split ratio is 80/20. For the training, we generate 256px x 256px crops of the original images using a sliding window. Min-max-normalization to a range [0,1] is applied to the raw data. Data augmentations (flipping, random crop & resize, brightness/contrast adjustment, rotate/shift/scale, Gaussian noise/blur) are used to extend the training dataset. Hyperparameters are obtained using random search.

The DNN is implemented in PyTorch Lightning. The image processing libraries OpenCV and scikit-image are used for noise simulation. DNN training is performed on multiple cluster nodes equipped with Intel Xeon Platinum 8368 CPU and four NVIDIA A100 Tensor Core GPU, respectively.

To avoid initialization effects, we repeat the experiments five times using various random seeds. Mean metrics including spread in form of standard deviation are shown.

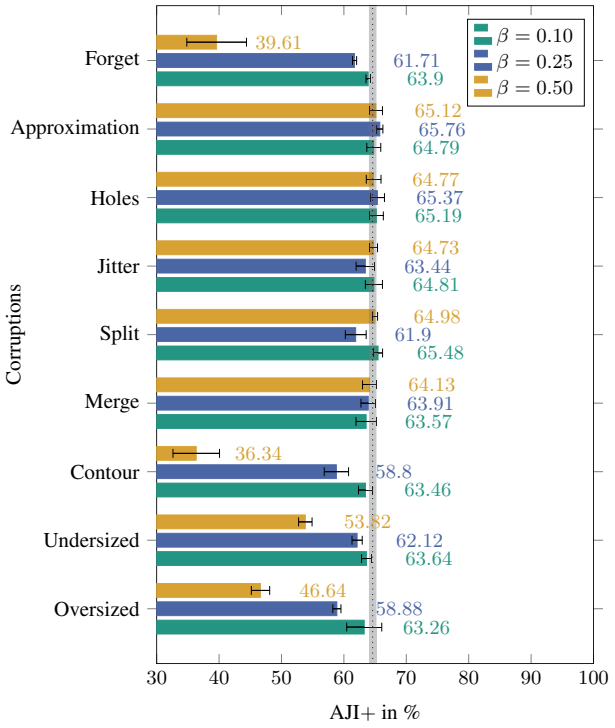


Fig. 4: Impact of Intra-Annotator Variability: Mean AJI+ (standard deviation visualized with black bars) w.r.t. $\mathcal{D}_{\text{test}}$ is given for different noisy datasets $\tilde{\mathcal{D}}_{\text{intra}}$ and corruption rates β. The baseline on the clean dataset $\mathcal{D}$ (AJI+ = 64.61±1.19 %) is depicted in gray.

3.3 Impact on Deep Learning

Intra-Annotator Variability

Figure 4 shows the impact of the introduced corruptions for different corruption rates $\beta = \{0.10, 0.25, 0.50\}$. The corruptions of "approximation", "jitter", and "holes" lead to no performance degradation and even in some cases to performance improvements. The AJI+ in the case of "split" or "merge" remains mostly in the baseline performance interval apart from a single outlier. However, the corruptions "forget", "contour", "undersized", and "oversized" show strong performance reduction which is related to the instance corruption ratio β. As this is a preliminary study, additional t-tests are not performed.

Inter-Annotator Variability

Simulating the division of labor case, the results of investigations in the context of inter-annotator variability are shown in Table 1. The combination of "approximation" corruption and no annotation noise leads to $\approx 0.5\%$ performance boost w.r.t. AJI+. However, all other combinations of corruptions yield a strong decreasing AJI+. In addition, the case of a partially clean dataset ("none") shows better performance measures compared to corrupting both parts of the dataset.

Tab. 1: Impact of Inter-Annotator Variability. Mean AJI+ ($\pm$ standard deviation) w.r.t. $\mathcal{D}_{\text{test}}$ is compared to noisy datasets $\tilde{\mathcal{D}}_{\text{inter}}$.

Corruptions		AJI+ in %
Annotator A	**Annotator B**	
None	None (Baseline)	64.61 $\pm$ 1.19
Oversized	Undersized	43.59 $\pm$ 2.10
Oversized	Contour	33.15 $\pm$ 4.26
Oversized	Approximation	45.37 $\pm$ 3.00
Oversized	None	45.51 $\pm$ 3.08
Undersized	Contour	5.53 $\pm$ 2.52
Undersized	Approximation	53.14 $\pm$ 3.83
Undersized	None	55.23 $\pm$ 3.68
Contour	Approximation	40.16 $\pm$ 13.78
Contour	None	32.35 $\pm$ 7.61
Approximation	None	65.08 $\pm$ 0.56

4 Discussion

Taking intra-annotator variability into account, the main message is that annotation noise does not generally lead to decreasing performance in instance segmentation. The most critical types of noise are forgetting instances, undersized or oversized instances, and contour corruption. Against expectation, merging/splitting instances seems no major issue if correct annotations dominate the total dataset. In addition, there might be performance improvements in the case of some minor corruptions, i.e., jitter, approximating the shape, or holes. On closer inspection, this appears to be a kind of annotation augmentation that may explain the better generalization of the DNN. Besides, this also means for experts that this kind of annotation noise poses no problem for the DNN training.

The results of inter-annotator variability show that different annotation styles impede the DNN training in mostly every case. Hence, in the division of labor cases, a consistent annotation policy between all annotators is of great significance. Less performance degradation in cases of only a partially corrupted dataset makes sense since the training can profit from partially correct labels. It should be noted, however, that contour corruption in the case of variability between annotators will not occur in practice to the extent simulated.

5 Conclusion

Annotator variability is already examined in classification and semantic segmentation. We categorized different forms of annotator variability in instance segmentation for the first time and proposed a concept to investigate its impact on DL pipelines. Minor annotation noise of a single annotator leads to no quantifiable performance drop. Hence, the accuracy requirements for annotators can be relaxed in this case. Nevertheless, the corruptions of forgetting instances, undersized/oversized instances, and missing filling of instances should be avoided to achieve high-quality DNNs. We demonstrated that different annotation styles are a major issue w.r.t. DNN performance. The agreement of annotation policies is necessary in division labor cases. Future work are investigations w.r.t. automated detection of noisy annotations and the potential of annotation augmentation to improve DL performance.

Acknowledgment: This work was supported in by the HoreKa Supercomputer through the Ministry of Science, Research, and the Arts Baden-Württemberg and by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

Author Statement

Research funding: This work was funded by the KIT Future Fields II Project "Screening Platform for Personalized Oncology". Conflict of interest: Authors state no conflict of interest.

References

[1] Juan C. Caicedo et al. “Nucleus Segmentation across Imaging Experiments: The 2018 Data Science Bowl.” In: *Nature Methods* 16.12 (2019), pp. 1247–1253.

[2] Simon Graham et al. “Hover-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images.” In: *Medical Image Analysis* (2019), p. 101563.

[3] Davood Karimi et al. “Deep Learning with Noisy Labels: Exploring Techniques and Remedies in Medical Image Analysis.” In: *Medical Image Analysis* 65.5 (2020), p. 101759.

[4] Neeraj Kumar et al. “A Multi-Organ Nucleus Segmentation Challenge.” In: *IEEE Transactions on Medical Imaging* 39.5 (2020), pp. 1380–1391.

[5] Katharina Löffler et al. “A Graph-Based Cell Tracking Algorithm with Few Manually Tunable Parameters and Automated Segmentation Error Correction.” In: *PLOS ONE* 16.9 (2021), pp. 1–28.

[6] Curtis Northcutt et al. “Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks.” In: *Neural Information Processing Systems Track on Datasets and Benchmarks*. Vol. 1. 2021.

[7] Olaf Ronneberger et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” In: *Medical Image Computing and Computer-Assisted Intervention*. Vol. 9351. 2015, pp. 234–241.

[8] Marcel P. Schilling et al. “Automated Annotator Variability Inspection for Biomedical Image Segmentation.” In: *IEEE Access* 10 (2022), pp. 2753–2765.

[9] Shaode Yu et al. “Robustness Study of Noisy Annotation in Deep Learning Based Medical Image Segmentation.” In: *Physics in Medicine & Biology* 65.17 (2020), p. 175007.