On the risk of manual annotations in 3D confocal microscopy image segmentation

Justin Sonneck, Shuo Zhao and Jianxu Chen
Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.
{justin.sonneck, shuo.zhao, jianxu.chen}@isas.de

1. Introduction

Nowadays, many influential biomedical researchers with a novel discovery in basic or translational science were enabled by modern microscopy techniques. For example, imaging over 0.2 million human induced pluripotent stem (hiPS) cells using high-resolution 3D confocal fluorescent microscopy built up the ground for the discovery of the fundamental principle of how normal stem cells organizing themselves [16]. With the advancement of microscopy techniques, the scale of microscopy images was also growing at an unprecedented pace, in terms of number of images or the size of each image, or both. Therefore, fully automated microscopy image analysis has become critical, where deep learning-based methods have achieved significant success in recent years.

In order to develop deep learning-based biomedical image segmentation models, manual annotations still play a decisive role, either as training data for supervised learning [6] or active learning [19], or as the ground truth for evaluating self-supervised methods [12], semi-supervised methods [21], or weakly supervised methods [18]. It is well-known that manual annotation could be extremely time-consuming especially in 3D, taking hours to annotate a single image, and also suffer from irreproducibility and subjectiveness issues, as it is impossible for people to annotate the same image pixel-wise identical when requested to do multiple times, or for different people to perform identical annotations on the same image.

Besides the time-demanding, irreproducible, and subjective concerns, the most important issue with manual annotation is the potential inaccuracy and therefore the potential negative impact on the downstream analysis or applications. For medical images, e.g., in radiology, it is possible that “a single source of truth” does not exist in practice and many methods have been proposed to leverage multi-annotations [20]. Even though we argue that the “real” truth can never be physically accessible, for certain bioimaging problems, it is feasible to obtain “a single source of truth” with high biological validity, which can be consid-

Abstract

Image segmentation in 3D confocal fluorescence microscopy images is a common problem in many biomedical studies. Deep learning-based methods have achieved great success on such tasks. In the literature, manual 3D annotations are still commonly used for model training or performance evaluation. But, due to the nature of the lens-based optical instruments, diffraction of light always occurs, which can lead to obscure boundaries of the biomedical structures being imaged. For example, when analyzing nuclei from 3D fluorescence microscopy images of cells marked by DNA dyes, the exact boundaries are usually not clearly identifiable, especially along Z. This makes accurate segmentation, both manually and automatically, very challenging. For applications where the boundary accuracy is crucial, the downstream analyses can thus be significantly compromised. This problem can be addressed with special experimental-computational co-design to acquire the “biological ground truth”. For the nuclei example, we can take cells expressing mEGFP tagged lamin B1, from which we can acquire both the DNA dye channel (nucleus) and the lamin B1 channel (nuclear envelope). Lamin B1 signals clearly mark the nuclei boundary and can thus serve as the real truth. We demonstrate that training a deep learning-based nuclei instance segmentation model with biological ground truth and manual annotations will result in significant differences in various metrics, such as volume or application-specific measurements. Also, we show the universality of such issues with manual annotations by testing different state-of-the-art deep learning-based methods. We hope our work can raise within the biomedical image analysis community the awareness of (1) the importance of interdisciplinary collaborations, e.g., computational-experimental co-design for biological ground truth collection, and (2) potentially significant issues with manual annotation in training or evaluating deep learning-based segmentation models.
erred as biologically optimal approximation of the real truth in practice. This is what we called “biological ground truth” (bioGT, see Section 2.2).

In this work, we used the problem of nuclei instance segmentation from high-resolution 3D fluorescent microscopy images, as an example to quantitatively demonstrate the risk of manual annotations. With a systematic experimental-computational co-design strategy, it is possible to acquire special data where the cells “inform” us a clear nuclear boundary by themselves as the bioGT. Evaluated by such bioGT, we find that the segmentation results obtained from a deep learning model trained with manual annotations could suffer from over 30% errors in the downstream analysis (see comprehensive analysis in Section 3). To the best of our knowledge, research has been done to urge the establishment of quality standard for manual annotations in medical images [10], and our work is the very first time where the risks of manual annotations in bioimages are systematically quantified. We hope to raise the awareness of the potential risks of manual annotations and at the same time highlight the interdisciplinary nature of biomedical image analysis, where experimental-computational co-design could be a viable path to achieve trustworthy deep learning-based biomedical image analysis.

2. Methods

The problem we used in this work to study the risk of manual annotation is 3D nuclei instance segmentation from high-resolution 3D confocal microscopy images of hiPS cells visualized via a fluorescence DNA dye. In this section, we will first present 2 different approaches for manual annotation, followed by the bioGT in this problem, then the deep learning models we used for 3D nuclei instance segmentation, and finally the evaluation metrics.

2.1. Manual Annotation

There are several approaches to manually annotate nuclei. One way is to annotate individual slices in XY projection of a 3-dimensional stack independently and then reassemble the obtained segmentation into a stack. We have done this using the Python-based multidimensional image viewer Napari [1] and call this annotation Napari ground truth (Napari-GT) hereafter. In contrast, we used the software 3D Slicer [4] to annotate multiple, but not all, XY-projected slices per nucleus and then interpolated between all segmented slices. A great strength of 3D Slicer is the combination of simultaneous display of front and side views. While annotating individual slices using 3D Slicer, we constantly considered all 3-axis combinations to annotate top and bottom of the nuclei as accurately as possible. We call the thus obtained annotation Slicer-GT. Fig. 2 gives insights into the corresponding graphical user interface, where front and both side views are simultaneously displayed. Indicated arrows show crosshairs for a better cross reference between different views.

2.2. Biological Ground Truth (bioGT)

With interdisciplinary experimental-computational co-design, it is possible to obtain biologically valid segmentation ground truth efficiently. In the high-resolution 3D confocal fluorescent microscopy image in this problem, the nuclei marked by DNA dye may suffer from very obscure boundaries (see Fig. 1-B), especially along the optical axis (a.k.a., Z-axis), due to light diffraction and photobleaching. As a result, it is very hard to delineate the nuclear boundary with pixel-level accuracy, particularly along Z. One option to obtain the bioGT, i.e., a biologically optimal approximation of the real truth, is as follows. With gene-editing techniques, one can make hiPS cells express-
2.3. Deep Learning-based 3D nuclei instance segmentation

In this work, we adopted a variant of EmbedSeg [7], a deep learning model with clustering-based instance generation, as the main method for 3D nuclei instance segmentation. The advantages of clustering-based instance segmentation models are, for example, independence of image dimensions and in theory able to handle instances of different connectivity and morphology. We used the extended EmbedSeg introduced in [14], which offers more effective training (e.g., on-the-fly data augmentation, training with exclusion masks, multi-GPU training, etc.) and more flexible network backbones (we chose the anisotropic variant of UNet introduced in [2]). In addition, to demonstrate the issue with models trained on manual annotations is universal, we tested another two state-of-the-art 3D nuclei instance segmentation methods: StarDist-3D [17], a deep learning-based method using star-convex polyhedra to represent nuclei (2D or 3D), and Cellpose [15], a method based on the prediction of spatial gradients. Here, we used the 3D extension provided by Cellpose, using 2D models to create 3D instance segmentation. A further development of Cellpose is Omnipose [3], which unlike the extension of Cellpose, uses 3D models for 3D segmentation. Since training of custom 3D models was not fully functioning at the time of the paper submission opening [5], nuclei instance segmentation using Omnipose is not pursued further in this manuscript. We didn’t include any detection-based instance segmentation [11], since densely packed squeezed 3D shapes pose a great challenge for accurately separating tightly touching boundaries with 3D bounding boxes.

2.4. Methodology of Validation

Pixel-level accuracy of the nuclei instance segmentation results may not matter for simple applications, like nuclei counting, but could be important for other downstream analysis. Besides basic metrics, like volume, we performed a systematic evaluation of the impact of the segmentation accuracy on the 3D nuclear shape modeling with spherical harmonic (SH) expansions. Using SH expansions, the nuclear shapes can be decomposed into individual parameters, like oscillations using the Fourier transform. The corresponding equation can be seen in equation (1), where $l$ is the SH degree, $m$ is the order, $\phi$ is the polar angle and $\theta$ is the azimuthal angle. $P_{lm}$ is the corresponding Legendre polynomial [13]. With all SH coefficients of the first 16 SH degrees, principal component analysis was used to identify a set of generally interpretable modes of the nuclear shapes [16]. We used the first 8 principal components (PCs) as used in [16] and 10 most important SH coefficients of each PC as the metrics for quantifying how the inaccuracy in manual annotations affect the downstream analysis.

$$Y_{lm}(\theta, \phi) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{2l + 1}{2} \frac{(l - m)!}{(l + m)!}} P_{lm}(\cos \theta) e^{im\phi}$$

3. Experiments, results and discussions

The source data in our experiments were from the public dataset released with [16]. We took a subset of multi-

---

Figure 2. Detail of the graphical user interface of the software 3D Slicer [4], which we used for Slicer-GT. We show front view and both side views including crosshairs (yellow arrows) for cross reference. In addition, a 3D representation of the annotated nucleus is shown in the upper right corner.
channel 100×3D microscopy images of hiPS cells with both the DNA-dye channel and the lamin B1 channel. First, we manually segmented 560 nuclei in 20 images, both Napari-GT and Slicer-GT, merely based on the DNA-dye channel without referring the lamin B1 channel. The manual annotations were done by PhD students with good knowledge of optical microscopy images. Meanwhile, we obtained the bioGT for the same set of images based on the lamin B1 channel using seeded watershed, as shown in Fig. 1-B. After collecting the manual annotations and the bioGT sets, we compared the volumes of the nuclei to study the reproducibility of the bioGT volumes using manual annotations. Based on these results, we used the Napari-GT and bioGT sets to train different state-of-the-art deep learning-based instance segmentation methods to further analyze how the choice of training data affects predicted segmentations. We held out an independent evaluation set of 18 images with bioGT. For the analysis we considered all predicted nuclei completely in the field of view (i.e., not touching image borders) with volume of at least 100 μm$^3$.

A nucleus in the prediction is considered as matched with a nucleus in the ground truth if there is an overlap of at least 51% of the volume of the ground truth nucleus. All computations were conducted on a computing cluster, only using one NVIDIA A100 GPU with 40GB memory. All the hyperparameters used in our experiments were according to the suggested values in the original packages, such as learning rate, optimizer, train/validation split ratio, etc. The scripts for downloading the raw microscopy images from the public data repository, running all the model trainings, evaluation and analysis will be released at https://github.com/MMV-Lab/manual_annotations_risk. The collected manual annotations and bioGT as well as the trained models will be available at https://zenodo.org/record/8247136, in order to make our results fully reproducible.

### 3.1. Comparison between manual annotations and bioGT

Fig. 4 shows a direct comparison between the volumes of individual nuclei in bioGT and both manual annotations (Slicer-GT and Napari-GT), with the coefficients of determination $r^2 = -0.81$ (Napari-GT) and $r^2 = -2.93$ (Slicer-GT). $r^2$ describes the goodness of a regression’s linear fit. The maximum value of 1 means perfect fit. In our case, higher $r^2$ indicates smaller discrepancy between the volumes of manual segmentation and bioGT. Negative values indicate a bad fit, where a simple average value is even better than the fit used. Specifically, the nuclear volume of manual Napari-GT segmentation is 32.41% ± 24.81% larger than the corresponding nucleus in bioGT. We believe the observed overestimation is mainly due to “optical illusion”, where the diffraction of light causes the objects appearing larger than they should be, especially along the optical axis. It is also evident, that Slicer-GT annotated nuclei have similar problems with overestimated volumes. Not only the coefficient of determination is even smaller, as indicated above, but also the deviation of the volumes is significantly higher significantly higher at 58.81% ± 22.06%.

We suspect that the constant side view of the software 3D Slicer might be misleading to detect the exact boundaries of DNA-dyed nuclei in Z. Additionally, interpolation between slices at the blurred boundaries could lead to inaccuracies. Due to the larger deviation of Slicer-GT annotated nuclei, we focused on Napari-GT as representative of manual annotation for further analysis. We can observe similar manual annotation issues in other public benchmark datasets, not only on our manual annotations. For example, in the dataset BBBC034v1 from the Broad Bioimage Benchmark Collection [8], the manual segmentation was performed by researchers with great expertise on these microscopy data. Fig. 3 shows an example of their nuclear segmentation, with reasonable accuracy in XY (even though slightly less smooth than it should be), while the segmentation along Z suffers over-segmentation issues. The annotation process of this data is not described in detail in the corresponding literature, but due to the inconsistencies visible in side view, a slice-by-slice annotation similar to Napari-GT can be assumed. We presented this example, published as ground truth data, to show how common such problems are for manual annotation.

![Image](Image.png)
3.2. Comparison between different deep learning models

On average, EmbedSeg, StarDist-3D and Cellpose take 1 ~ 9 minutes (depending on pre-training with pre-cropped patches or fine-tuning on full images), 2 minutes, and 7.5 minutes for each training epoch, as well as 21GB, 39GB, 4.7GB GPU memory, respectively. For segmentation performance, out of 292 nucleus instances in total, EmbedSeg and StarDist-3D correctly detected 291 and 292, respectively, while Cellpose only detected 98, suffering from many false negative errors. This is probably due to Cellpose’s 2D-based pseudo-3D training. Next, we compared the volume of segmented nuclei from each model against the volume of the nuclei in bioGT as a ratio: \((\frac{Vol_{seg}}{Vol_{bioGT}} - 1) \times 100\%\). We got EmbedSeg\textsubscript{bioGT} (4.04\% \pm 3.70\%), EmbedSeg\textsubscript{Napari-GT} (29.58\% \pm 12.33\%), StarDist-3D\textsubscript{bioGT} (15.24\% \pm 8.07\%), StarDist-3D\textsubscript{Napari-GT} (17.53\% \pm 13.36\%), Cellpose\textsubscript{bioGT} (12.62\% \pm 8.75\%), Cellpose\textsubscript{Napari-GT} (23.74\% \pm 16.99\%), respectively. We can see that the EmbedSeg model trained with manual Napari-GT annotation resulted in about 30% over-estimation in nuclear volume, while only < 5\% for the EmbedSeg model trained with bioGT.

Table 1 summarizes the \(r^2\) values measuring how consistent every metric was between segmentation from different models and bioGT. EmbedSeg\textsubscript{bioGT} achieved the best overall performance, much better than EmbedSeg\textsubscript{Napari-GT}. Models trained with bioGT generally performed better than models trained with manual annotations, even though the impact varies for different models. StarDist-3D\textsubscript{bioGT} performed only slightly better than StarDist-3D\textsubscript{Napari-GT} while both suffering from considerable inaccuracy. This is mainly
due to the underlying star-convexity, making segmented nuclear boundaries more polygonal and less accurate. An example can be seen in Fig. 5B. It can be seen that the boundaries of the two predictions in XY at the arrow-indicated areas differ significantly from bioGT. In addition, it is noticeable that StarDist-3D\textsubscript{Napari-GT} has clear problems in determining the correct nucleus height. A direct comparison between EmbedSeg\textsubscript{bioGT} and StarDist-3D\textsubscript{bioGT} can be seen in Fig. 5A. It can be seen that the segmentation of the EmbedSeg\textsubscript{bioGT} model matches the exact boundary of the nucleus much better. To emphasize the blurry boundaries, the contour of the bioGT is also shown in the DNA dye representation. The superior performance of Cellpose\textsubscript{bioGT} over Cellpose\textsubscript{Napari-GT} is likely because bioGT has better spatial consistency in 3D and thus much easier for Cellpose to learn spatial gradients (i.e., annotating 3D nuclei slice-by-slice can hardly maintain the smoothness along Z). Fig. 5C shows an example of both Cellpose predictions. Both cases clearly show that Cellpose has problems to segment the nuclei properly. Not only are the boundaries predicted to be too small, in addition many different cell instances are predicted, thus splitting a single nucleus into many smaller segments. However, Cellpose\textsubscript{bioGT} predicts a clearly dominant instance, which is why the volume of Cellpose\textsubscript{bioGT} is closer to that of bioGT than that of Cellpose\textsubscript{Napari-GT}.

Figure 6. Overview of the difference between the coefficients of determination \(\Delta r^2\) of the predictions obtained with EmbedSeg\textsubscript{bioGT} and EmbedSeg\textsubscript{Napari-GT}. Positive values indicate the superiority of the bioGT model, while the annotated cells of the heatmap represent the coefficients discussed with high impact on the principal components presented. Accordingly, the other coefficients shown are negligible due to their low impact on the principle components and are shown only for completeness. For clarity, the shown range is limited to ±0.5.
3.3. Impact on individual SH coefficients

Following the first 8 PCs (covering 86.33% of the variance within the data), we further examined 10 SH coefficients with the largest impact on each of these 8 PCs. Since one SH coefficient may be important for multiple PCs, in our case, this resulted in 35 different SH coefficients in total as metrics, which we compared the difference between their \( r^2 \) for EmbedSeg\textsubscript{bioGT} and EmbedSeg\textsubscript{Napari-GT}. The resulting error is \( \Delta r^2 \) in a range of \( 0.08 \leq \Delta r^2 \leq 2.61 \) with an average of 0.43 (positive mean EmbedSeg\textsubscript{bioGT} has higher \( r^2 \), while negative mean EmbedSeg\textsubscript{Napari-GT} has higher \( r^2 \)). Fig. 6 shows the comparison of all \( \Delta r^2 \). Even though only a total of 50.52% of the coefficients are superior for the \textit{bioGT} model, this includes, among others, all 35 components with a large impact on our analysis. The first principal component covers 45.51 % of the total variance within the data and the average absolute weight \( \omega \) of the top 10 coefficients for PC1 is \( \omega = 0.211 \pm 0.235 \) in a range of \( 0.037 \leq \omega \leq 0.807 \), while the average absolute weight of the other 279 coefficients (including 25 ones with large impact on other PCs) for the first PC is \( \omega = 0.002 \pm 0.003 \) in a range of \( 0.000 \leq \omega \leq 0.023 \). This shows the negligibility of the coefficients without much impact on the PCs and, since the \textit{bioGT} model is superior for all impactful SH coefficients, this is further evidence of how manual annotations affect downstream modeling of nuclear shapes.

3.4. Impact of the size of training data

We conducted experiments on different amounts of training data for EmbedSeg using \textit{bioGT}. The results are shown in Fig. 7A and confirm, based to the improved performance with an increased number of training images, that a sufficiently large amount of training data is important. The size of training data heavily depends on the difficulty and time-cost of ground truth collection. For a 3D image of about 30 nuclei, the manual \textit{Napari-GT} annotation time was about one hour, while the total human effort to generate the \textit{bioGT}, at less than a minute, was a fraction of that time. Therefore, a systematic experimental-computational co-design to permit efficient \textit{bioGT} collection would make it painless to collect a fairly large training set and thus possible to build robust deep learning models for subsequent analyses. In addition, Fig. 7B shows the only failed case, where the prediction of the EmbedSeg\textsubscript{bioGT} model contains holes within the segmentation, while the prediction of EmbedSeg\textsubscript{Napari-GT} also segments only slightly more than half of the shown nucleus. This is a special nucleus where the DNA dye was not highly expressed, resulting in very weak nuclear signals without much textures. This is another indication of the need for sufficient training data to cover enough variance and make the predictions more robust against heterogeneously distributed DNA dye.

4. Conclusions and outlook

In this work, we want to point out the non-trivial risk of using manual segmentation in biomedical research with quantitative evidence, using an example problem of 3D nuclei instance segmentation in fluorescence confocal microscopy. We showed the universality of this issue on various state-of-the-art deep learning-based 3D nuclei instance segmentation models and quantified the need for a sufficiently large training set, which also made manual annotation practically infeasible due to the huge time requirement. Based on our results, we propose a computational and experimental co-design for the collection of biological ground truth data that can be used to train deep learning-based seg-
mentation models to enable robust and reproducible analysis, even in subsequent experiments in which no additional channel for bioGT is included anymore. In experiments where obtaining a bioGT is not possible, we propose to perform variance estimation where a sample of the dataset is annotated multiple times by the same researcher or the same sample is annotated independently by multiple experts and these annotations are evaluated for robustness.

Subsequent studies could address the detailed examination of data acquisition of further biologically correct ground truths for other biomedical data and discuss their limitations and advantages. It would also be interesting to examine the impact of (an)isotropic sampling in Z as well as that of convolved fluorescence light on ground truth data acquisition. Possible further subsequent work could be the investigation of specific applications: Although for certain experiments, knowing the volume is crucial, in some cases it is already sufficient to know it only relatively (for example, before and after treating the sample with a drug to analyze the drug’s impact on growth or shrinkage). For these experiments, studies could be conducted to determine whether deep learning-based instance segmentation methods trained with manual annotations can predict volume changes consistently enough for reliable studies.

We hope that our work has brought enough attention to the biomedical image analysis community regarding the potential issues of manual segmentation and highlights the necessity of interdisciplinary collaboration for reliable, time-efficient and reproducible biomedical image analysis.

Acknowledgement

All authors are funded by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) in Germany under the funding reference 161100272 and are also supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia (Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen, MKW NRW).

References


