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

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 universalness 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.

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-

ered 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.

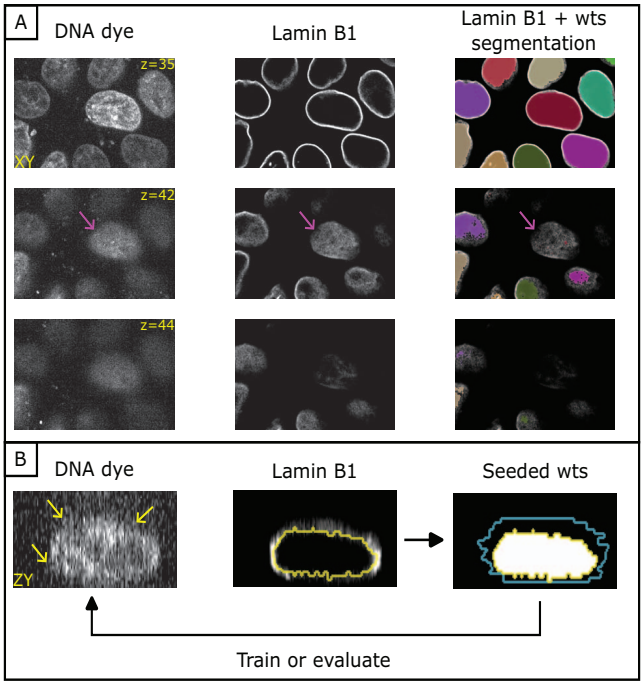

Figure 1. (A) Different Z slices (same XY section) of a 3D image with DNA dye channel, lamin B1 channel and its overlay with the obtained watershed segmentation. The arrows in magenta indicate some signals coming from light diffraction near the top of a cell, which should not be segmented. (B) Overview of the workflow for obtaining *bioGT* with seeded watershed segmentation in the lamin B1 channel. The watershed result (white areas in the bottom image), whose contour (in yellow) accurately follows the lamin B1 signal (the middle image), can be used as the *bioGT* (i.e., the segmentation target) for nucleus segmentation in the DNA dye channel (the top image), where the nuclear boundary is very unclear (see yellow arrows). As a reference, we overlay the contour of manual segmentation (*Napari-GT*, in blue) with the *bioGT*. The over-segmentation by manual annotation is mainly due to the blurry boundary in the DNA channel, especially along Z direction.

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-

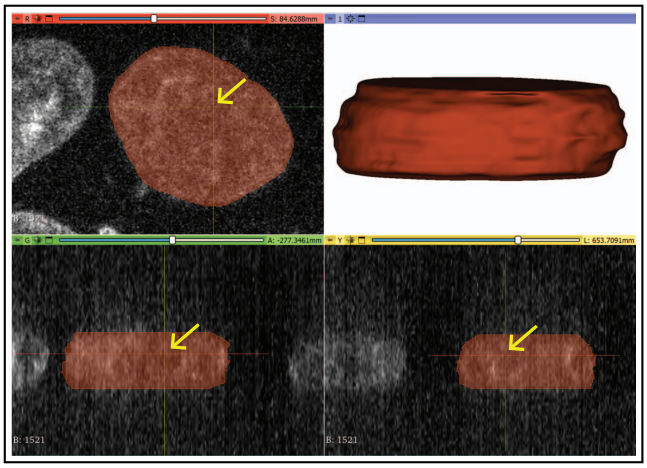

Figure 2. Detail of the graphical user interface of the software *3D Slicer*[4], which we used for *Slicer-GT*. We show front view and both side views including crosshairs (yellow arrows) for cross reference. In addition, a 3D representation of the annotated nucleus is shown in the upper right corner.

ing mEGFP tagged lamin B1, and design an experimental assay to acquire multi-channel 3D images, with both the DNA dye channel and the endogenously tagged lamin B1 channel (see Fig. 1A). Lamin B1 is a protein that marks up the nuclear lamina, commonly used for visualizing the nuclear envelope, which could be generally interpreted as the "membrane" of the nucleus. The light diffraction effect on such thin membrane-like signals is negligible comparing to the ball-like signals of nuclei. Therefore, localizing the contours from the lamin B1 channel could provide a biologically valid approximate of the true nuclear boundary, and thus could be then used as the *bioGT*. There are many viable solutions to extract the nuclear shape from the lamin B1 channel. For simplicity, we used the seeded watershed algorithm [9], a semi-automatic method with human clicking inside the contours as seeds (taking less than a minute), and applied it to the lamin B1 channel itself. The results of the seeded watershed algorithm were not sensitive to the exact location where human clicked, and thus making the results well reproducible. It is worth mentioning that the lamin B1 channel is not always available; otherwise, we can simply segment nuclei directly from lamin B1. For other cell lines, e.g, gene-edited to express other proteins to visualize other cellular structures, we can only acquire the DNA dye channel, but not the lamin B1 channel.

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, ϕ is the polar angle and θ 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.

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

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-

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 $100\,\mu\mathrm{m}^3$. 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 computation was 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 $32.41\% \pm 24.81\%$ 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 $58.81\% \pm 22.06\%$. 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.

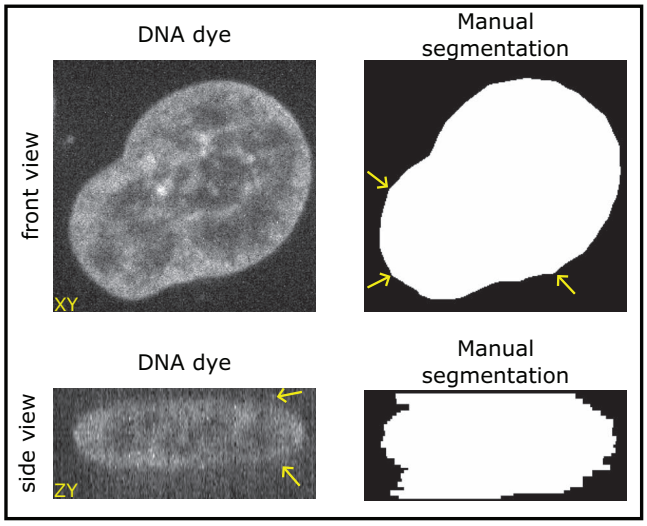

Figure 3. Overview of a manual annotation example including DNA dye channel and the manual segmentation for both front view and side view. It is worth mentioning that this is from a public data repository [8], not our annotation.

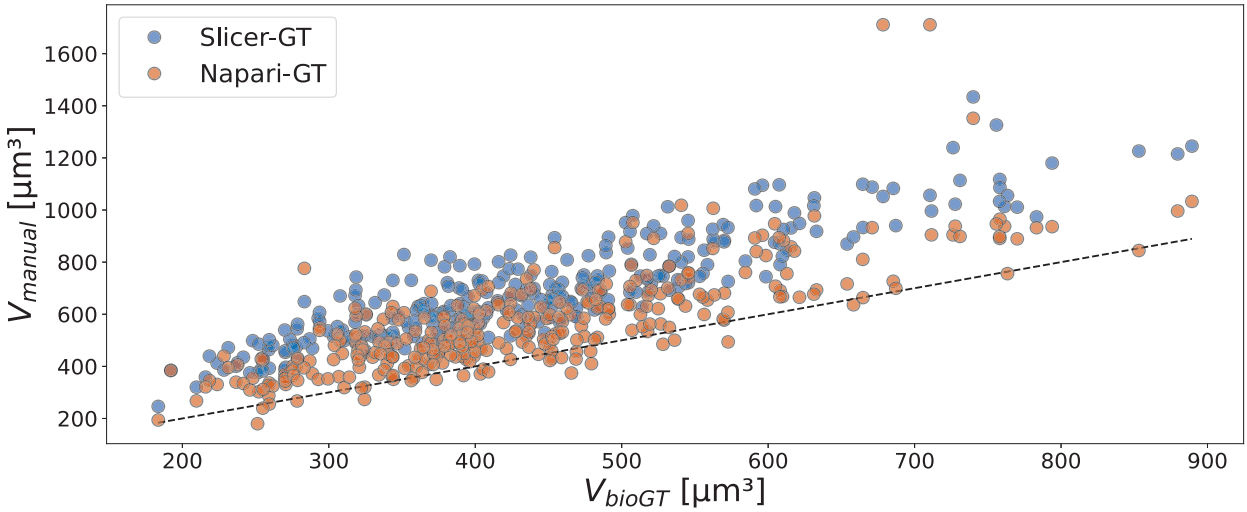

Figure 4. Comparison of nuclear volumes of both manual annotations and biological ground truth (*bioGT*). The unit line is plotted as a reference.

Table 1. r^2 for various metrics for different deep learning models, each trained with manual *Napari-GT* annotations and *bioGT*.

Metrics	EmbedSeg		StarDist-3D		Cellpose	
	Napari-GT	*bioGT*	*Napari-GT*	*bioGT*	*Napari-GT*	*bioGT*
Volume	-0.29	0.95	0.50	0.49	-0.10	0.70
PC1	0.38	0.96	0.45	0.94	-0.22	0.15
PC2	0.26	0.94	0.61	0.59	0.19	0.68
PC3	0.75	0.94	0.45	0.66	0.53	0.91
PC4	0.83	0.96	0.54	0.52	0.71	0.78
PC5	0.75	0.95	0.36	0.66	0.45	0.82
PC6	0.60	0.90	0.23	0.40	-0.49	0.58
PC7	0.25	0.77	0.24	0.44	-0.88	0.13
PC8	0.71	0.89	0.36	0.41	-0.06	0.65

3.2. Comparison between different deep learning models

On average, EmbedSeg, StarDist-3D and Cellpose take $1 \sim 9$ 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: $((Vol_{seg}/Vol_{bioGT} - 1) \times 100\%)$. We got $\text{EmbedSeg}_{\text{bioGT}}$ $(4.04\% \pm 3.70\%)$, $\text{EmbedSeg}_{\text{Napari-GT}}$ $(29.58\% \pm 12.33\%)$, $\text{StarDist-3D}_{\text{bioGT}}$ $(15.24\% \pm 8.07\%)$, $\text{StarDist-3D}_{\text{Napari-GT}}$ $(17.53\% \pm 13.36\%)$, $\text{Cellpose}_{\text{bioGT}}$ $(12.62\% \pm 8.75\%)$, $\text{Cellpose}_{\text{Napari-GT}}$ $(23.74\% \pm 16.99\%)$, 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*. $\text{EmbedSeg}_{\text{bioGT}}$ achieved the best overall performance, much better than $\text{EmbedSeg}_{\text{Napari-GT}}$. Models trained with *bioGT* generally performed better than models trained with manual annotations, even though the impact varies for different models. $\text{StarDist-3D}_{\text{bioGT}}$ performed only slightly better than $\text{StarDist-3D}_{\text{Napari-GT}}$ while both suffering from considerable inaccuracy. This is mainly

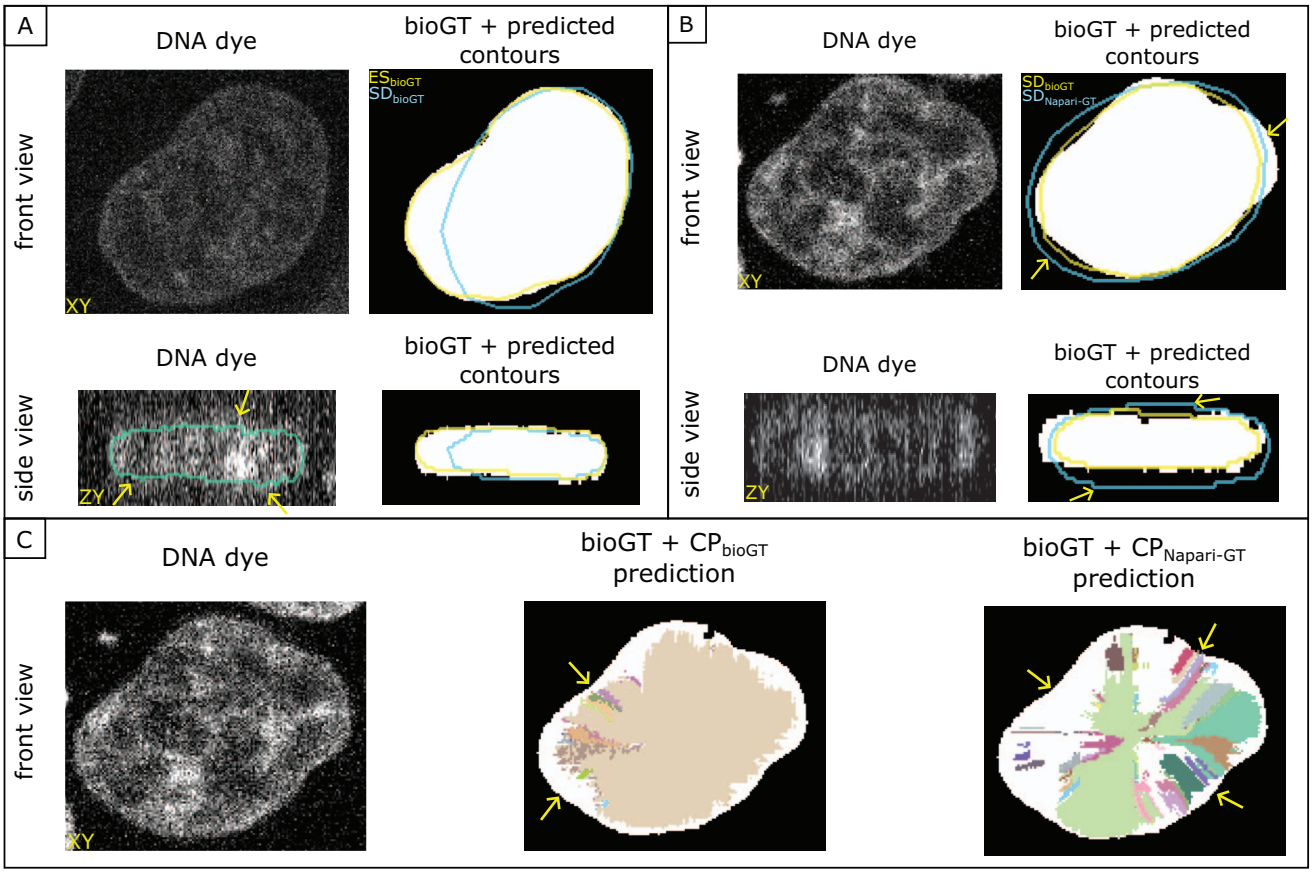

Figure 5. Sample prediction from (A) EmbedSeg and StarDist-3D, both trained with *bioGT*, (B) StarDist-3D$_{bioGT}$ and StarDist-3D$_{Napari-GT}$ and (C) Cellpose$_{bioGT}$ and Cellpose$_{Napari-GT}$, either in front view or in front and side view. White areas show the corresponding *bioGT*.

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$_{Napari-GT}$ has clear problems in determining the correct nucleus height. A direct comparison between EmbedSeg$_{bioGT}$ and StarDist-3d$_{bioGT}$ can be seen in Fig. 5A. It can be seen that the segmentation of the EmbedSeg$_{bioGT}$ 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$_{bioGT}$ over Cellpose$_{Napari-GT}$ 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$_{bioGT}$ predicts a clearly dominant instance, which is why the volume of Cellpose$_{bioGT}$ is closer to that of *bioGT* than that of Cellpose$_{Napari-GT}$.

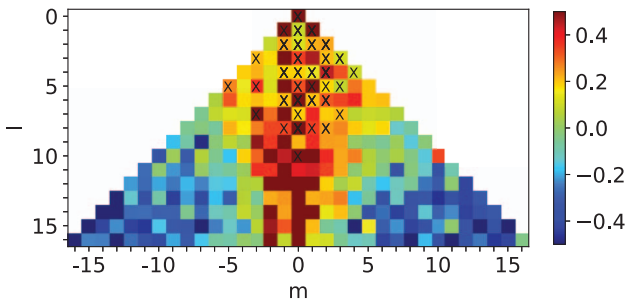

Figure 6. Overview of the difference between the coefficients of determination Δr^2 of the predictions obtained with EmbedSeg$_{bioGT}$ and EmbedSeg$_{Napari-GT}$. Positive values indicate the superiority of the *bioGT* model, while the annotated cells of the heatmap represent the coefficients discussed 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.

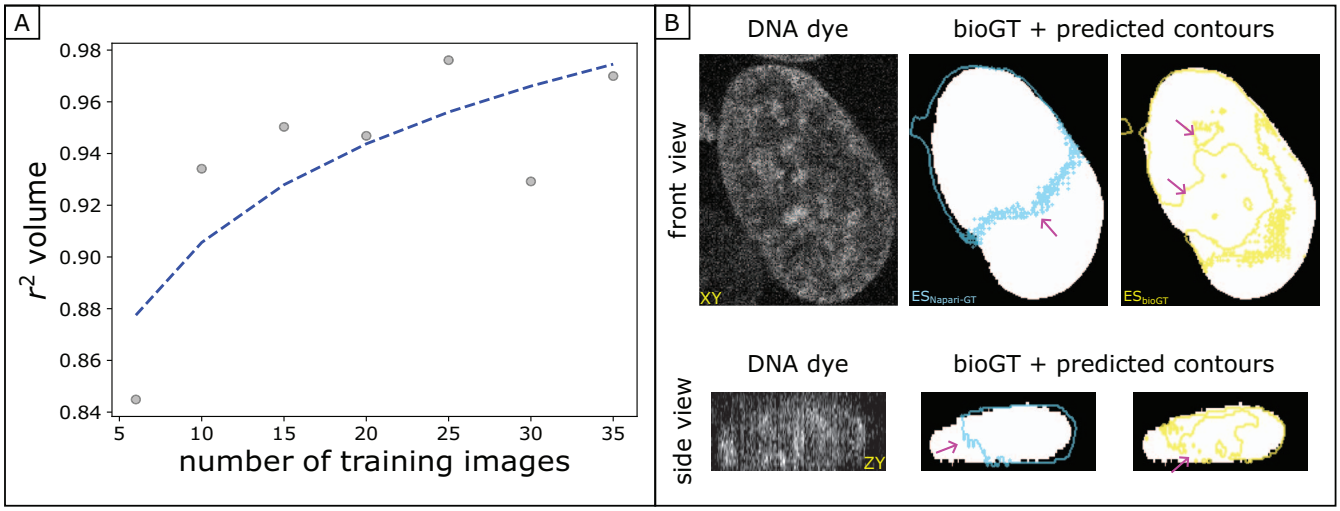

Figure 7. (A) r^2 of different EmbedSeg$_{\text{bioGT}}$ models with different numbers of training data. (B) Prediction from EmbedSeg$_{\text{bioGT}}$ and EmbedSeg$_{\text{Napari-GT}}$ on the only failed case (EmbedSeg detected 291 out of 292 nuclei), both compared to *bioGT* (the white area) in front overview and side view.

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$_{\text{bioGT}}$ and EmbedSeg$_{\text{Napari-GT}}$. The resulting error is Δr^2 in a range of $0.08 \leq \Delta r^2 \leq 2.61$ with an average of 0.43 (positive mean EmbedSeg$_{\text{bioGT}}$ has higher r^2, while negative mean EmbedSeg$_{\text{Napari-GT}}$ has higher r^2). Fig. 6 shows the comparison of all Δr^2. Even though only a total of $50.52\,\%$ of the coefficients are superior for the *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 $\bar{\omega}$ of the top 10 coefficients for PC1 is $\bar{\omega} = 0.211 \pm 0.235$ 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 $\bar{\omega} = 0.002 \pm 0.003$ 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 *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 *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 *Napari-GT* annotation time was about one hour, while the total human effort to generate the *bioGT*, at less than a minute, was a fraction of that time. Therefore, a systematic experimental-computational co-design to permit efficient *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$_{\text{bioGT}}$ model contains holes within the segmentation, while the prediction of EmbedSeg$_{\text{Napari-GT}}$ 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 universalness 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 161L0272 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

[1] Jannis Ahlers, Daniel Althviz Moré, Oren Amsalem, Ashley Anderson, Grzegorz Bokota, Peter Boone, Jordão Bragantini, Genevieve Buckley, Alister Burt, Matthias Bussonnier, Ahmet Can Solak, Clément Caporal, Draga Doncila Pop, Kira Evans, Jeremy Freeman, Lorenzo Gaifas, Christoph Gohlke, Kabilar Gunalan, Hagai Har-Gil, Mark Harfouche, Kyle I. S. Harrington, Volker Hilsenstein, Katherine Hutchings, Talley Lambert, Jessy Lauer, Gregor Lichtner, Ziyang Liu, Lucy Liu, Alan Lowe, Luca Marconato, Sean Martin, Abigail McGovern, Lukasz Migas, Nadalyn Miller, Hector Muñoz, Jan-Hendrik Müller, Christopher Nauroth-Kreß, Juan Nunez-Iglesias, Constantin Pape, Kim Pevey, Gonzalo Peña-Castellanos, Andrea Pierré, Jaime Rodríguez-Guerra, David Ross, Loic Royer, Craig T. Russell, Gabriel Selzer, Paul Smith, Peter Sobolewski, Konstantin Sofiiuk, Nicholas Sofroniew, David Stansby, Andrew Sweet, Wouter-Michiel Vierdag, Pam Wadhwa, Melissa Weber Mendonça, Jonas Windhager, Philip Winston, and Kevin Yamauchi. napari: a multi-dimensional image viewer for Python.

[2] Jianxu Chen, Liya Ding, Matheus P. Viana, HyeonWoo Lee, M. Filip Sluezwski, Benjamin Morris, Melissa C. Hendershott, Ruian Yang, Irina A. Mueller, and Susanne M. Rafelski. The allen cell and structure segmenter: a new open source toolkit for segmenting 3d intracellular structures in fluorescence microscopy images. *bioRxiv*, 2020.

[3] Kevin J. Cutler, Carsen Stringer, Teresa W. Lo, Luca Rappez, Nicholas Stroustrup, S. Brook Peterson, Paul A. Wiggins, and Joseph D. Mougous. Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. *Nature Methods*, 19(11):1438–1448, Nov 2022.

[4] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V Miller, Steve Pieper, and Ron Kikinis. 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging*, 30(9):1323–1341, Nov. 2012.

[5] GitHub Community. Omnipose Issue #28. `https://github.com/kevinjohncutler/omnipose/issues/28`, 2022. Online; accessed: June 21, 2023.

[6] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Feb 2021.

[7] Manan Lalit, Pavel Tomancak, and Florian Jug. Embedseg: Embedding-based instance segmentation for biomedical microscopy data. *Medical Image Analysis*, 81:102523, 2022.

[8] Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7):637–637, Jul 2012.

[9] F. Meyer and S. Beucher. Morphological segmentation. *Journal of Visual Communication and Image Representation*, 1(1):21–46, 1990.

[10] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Schreck, A. Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. Labelling instructions matter in biomedical image analysis. *Nature Machine Intelligence*, Mar 2023.

[11] Gregor N. Ramien, Paul F. Jaeger, Simon A. A. Kohl, and Klaus H. Maier-Hein. Reg r-cnn: Lesion detection and grading under noisy labels. In Hayit Greenspan, Ryutaro Tanno, Marius Erdt, Tal Arbel, Christian Baumgartner, Adrian Dalca, Carole H. Sudre, William M. Wells, Klaus Drechsler, Marius George Linguraru, Cristina Oyarzun Laura, Raj Shekhar, Stefan Wesarg, and Miguel Ángel González Ballester, editors, *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*, pages 33–41, Cham, 2019. Springer International Publishing.

[12] Michael C. Robitaille, Jeff M. Byers, Joseph A. Christodoulides, and Marc P. Raphael. Self-supervised

machine learning for live cell imagery segmentation. *Communications Biology*, 5(1):1162, Nov 2022.

[13] Xiongtao Ruan and Robert F Murphy. Evaluation of methods for generative modeling of cell and nuclear shape. *Bioinformatics*, 35(14):2475–2485, 12 2018.

[14] Justin Sonneck and Jianxu Chen. Mmv_im2im: An open source microscopy machine vision toolbox for image-to-image transformation. *arXiv*, 2022.

[15] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, Jan 2021.

[16] Matheus P. Viana, Jianxu Chen, Theo A. Knijnenburg, Ritvik Vasan, Calysta Yan, Joy E. Arakaki, Matte Bailey, Ben Berry, Antoine Borensztejn, Eva M. Brown, Sara Carlson, Julie A. Cass, Basudev Chaudhuri, Kimberly R. Cordes Metzler, Mackenzie E. Coston, Zach J. Crabtree, Steve Davidson, Colette M. DeLizo, Shailja Dhaka, Stephanie Q. Dinh, Thao P. Do, Justin Domingus, Rory M. Donovan-Maiye, Alexandra J. Ferrante, Tyler J. Foster, Christopher L. Frick, Griffin Fujioka, Margaret A. Fuqua, Jamie L. Gehring, Kaytlyn A. Gerbin, Tanya Grancharova, Benjamin W. Gregor, Lisa J. Harrylock, Amanda Haupt, Melissa C. Hendershott, Caroline Hookway, Alan R. Horwitz, H. Christopher Hughes, Eric J. Isaac, Gregory R. Johnson, Brian Kim, Andrew N. Leonard, Winnie W. Leung, Jordan J. Lucas, Susan A. Ludmann, Blair M. Lyons, Haseeb Malik, Ryan McGregor, Gabe E. Medrash, Sean L. Meharry, Kevin Mitcham, Irina A. Mueller, Timothy L. Murphy-Stevens, Aditya Nath, Angelique M. Nelson, Sandra A. Oluoch, Luana Paleologu, T. Alexander Popiel, Megan M. Riel-Mehan, Brock Roberts, Lisa M. Schaefbauer, Magdalena Schwarzl, Jamie Sherman, Sylvain Slaton, M. Filip Sluzewski, Jacqueline E. Smith, Youngmee Sul, Madison J. Swain-Bowden, W. Joyce Tang, Derek J. Thirstrup, Daniel M. Toloudis, Andrew P. Tucker, Veronica Valencia, Winfried Wiegraebe, Thushara Wijeratna, Ruian Yang, Rebecca J. Zaunbrecher, Ramon Lorenzo D. Labitigan, Adrian L. Sanborn, Graham T. Johnson, Ruwanthi N. Gunawardane, Nathalie Gaudreault, Julie A. Theriot, and Susanne M. Rafelski. Integrated intracellular organization and its variations in human ips cells. *Nature*, 613(7943):345–354, Jan 2023.

[17] Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3655–3662, 2020.

[18] Udaranga Wickramasinghe, Patrick Jensen, Mian Shah, Jiancheng Yang, and Pascal Fua. Weakly supervised volumetric image segmentation with deformed templates. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 422–432, Cham, 2022. Springer Nature Switzerland.

[19] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pages 399–407, Cham, 2017. Springer International Publishing.

[20] Dewen Zeng, Mingqi Li, Yukun Ding, Xiaowei Xu, Qiu Xie, Ruixue Xu, Hongwen Fei, Meiping Huang, Jian Zhuang, and Yiyu Shi. Segmentation with multiple acceptable annotations: A case study of myocardial segmentation in contrast echocardiography. *CoRR*, abs/2106.15597, 2021.

[21] Hao Zheng, Lin Yang, Jianxu Chen, Jun Han, Yizhe Zhang, Peixian Liang, Zhuo Zhao, Chaoli Wang, and Danny Chen. Biomedical image segmentation via representative annotation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:5901–5908, 07 2019.