AI-assisted labeling and its pitfalls: A case study in Electron Microscopy segmentation

Christina Bukas[1][0000-0001-9913-8525], Jonas Engler[2][0009-0003-4233-0868], Helena Pelin[1], Jonas Hagenberg[1][0000-0002-1849-1106], Haider Khan[1][0009-0001-7529-7086], Francesco Campi[1][0009-0008-2404-8582], Carola Eberhagen[3], Hans Zischka[2,3][0000-0002-4047-1566], and Marie Piraud[1][0000-0002-4917-2458]

[1] Helmholtz AI, Computational Health Center (CHC), Helmholtz Munich, Neuherberg, Germany
[2] Institute of Toxicology and Environmental Hygiene, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany
[3] Institute of Molecular Toxicology and Pharmacology, Helmholtz Munich, German Research Center for Environmental Health, Neuherberg, Germany

Abstract. High-quality labels are of utmost importance in biomedicine and instrumental for tasks such as disease understanding, drug discovery, and medical diagnosis. Human-in-the-loop approaches and AI-assisted annotations are currently widely accepted as the safest approach for data curation, offering both speed and human supervision, while leveraging the power of AI. Indeed, foundation models are often used to obtain initial labels which are then manually corrected by a human expert and subsequently used for downstream tasks. In this work, we uncover a rarely discussed risk of such semi-automated approaches and present a case study to demonstrate this. Focusing on microscopy imaging segmentation, where annotation quality is critical and costly, we collected and annotated a Transmission Electron Microscopy dataset in a semi-automated approach, using BATS [4], a software we built for AI-assisted segmentation labeling in microscopy which encourages data centric practices. We show how batch effects can be derived by a mix of human and model annotations affecting the downstream analysis. Finally, we demonstrate a solution which mitigates these batch effects and discuss how future researchers can adopt responsible and accurate data annotation pipelines.

Keywords: AI-assisted annotation · microscopy segmentation · batch effect

[4] https://github.com/HelmholtzAI-Consultants-Munich/BATS

1 Introduction

It goes without saying that few insights can be drawn from imaging data without high-quality annotations to accompany them [20,27]. In biomedical contexts, the need for carefully curated data becomes even more pressing, as mislabeled data can lead to anything from a dire misdiagnosis in medical diagnostics, to missing groundbreaking phenotypic relations in biological research [1]. In the current era of AI, much time can be saved during annotation with the use of automated tools, incorporating machine learning models for imaging tasks such as classification, object detection, or segmentation [3, 19]. In addition to time spared, human biases can be mitigated with the use of robust, reproducible pipelines, aspects which become increasingly important when considering big data required for training large foundation models [13, 17, 28]. However, one needs to take care when employing such models, which can regularly suffer from issues, such as failure to generalize well to out-of-distribution data, inherent biases etc. [18]. The use of human-in-the-loop (HITL) strategies can offer the best of both worlds: the speed and consistency of a machine, along with the expertise and general oversight of a human [11]. With this in mind, a desirable pipeline for well-curated data, typically involves a foundation model for obtaining an initial round of labels, which are then manually corrected and used to either fine-tune the model or a subsequent downstream task. This approach seems promising and is indeed being fast-adopted by biologists and clinicians alike.

However, in this work, we caution scientists against accepting this practice as the golden standard and show how semi-automated datasets can carry batch effects, affecting subsequent downstream tasks, e.g. model training or data analysis (Fig. 1; left). To demonstrate the pitfalls involved we designed a case study: we collected and curated a Transmission Electron Microscopy dataset of healthy and diseased mitochondria, the latter affected by Wilson's disease. Additionally, we present BATS, a Bioimage Annotation Tool for Segmentation. BATS is a software designed for AI-assisted microscopy image annotation while following data centric practices. Here, we chose to tackle the task of segmentation since acquiring such labels enables a wide range of downstream tasks. It is also one of the most laborious annotation tasks, therefore strongly benefiting from AI-assistance. Moreover, segmentation masks can easily be converted to bounding boxes or image-level labels, while the opposite does not hold true. We focus on microscopy imaging data as, again, here we see the great potential for AI-assisted approaches, since these data typically involve many instances whose detection is not trivial. We used our tool to annotate the dataset in a semi-automated setup and exported meaningful mitochondrial shape features from the acquired labels. We analyzed these features and show how a mix of model and human-edited labels can cause batch effects in the downstream data, thus affecting follow-up tasks. We demonstrate this by fine-tuning a segmentation model on the curated data, focusing on mitochondrial borders where human and model annotations differ in style. Finally, we propose a solution to mitigate these effects and ensure responsible data annotation (Fig. 1; right).

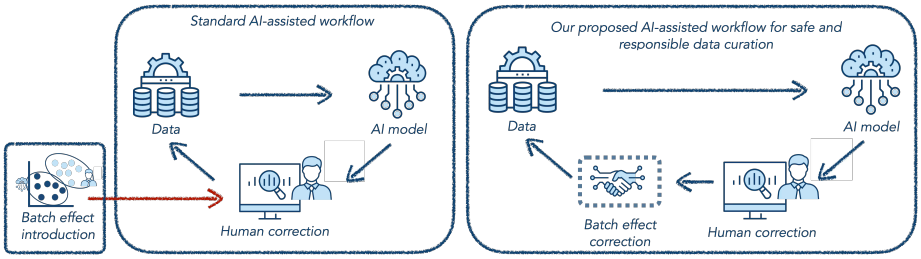

Fig. 1. Typical AI-assisted approaches ignore the batch effects introduced by mixing human and model labels. Our proposed AI-assisted pipeline corrects these effects.

2 Related work

While HITL approaches are steadily gaining popularity, little research exists on the potential risks of mixed label datasets. Kumar et al. discuss challenges of HITL approaches [15], describing several factors affecting performance, such as human factors, cost, and interpretability. In [21], the authors discuss human error describing how the order in which data is presented to human annotators in HITL systems affects data quality In both works, the aspects of mixing human and automated labels are not discussed. In the field of AI-assisted labeling much work focuses on faster and more accurate annotation. Radeta et al. conduct a controlled study of real-time video annotation with AI-assisted expert and non-expert annotators [24]. Their results show that AI assistance improves annotation quality, particularly for non-experts, and that models trained on labels combining AI and expert annotations achieve the best performance compared to purely human-annotated data. Similarly, Garcia-Sineriz et al. show how training a model with a dataset derived from an AI-assisted setup increases its performance, again comparing pure human-generated labels with human and AI-assisted labels [10]. Similar works propose workflows and tools for AI-assisted labeling, to boost performance and save valuable time [6,22,29]. However, to the best of our knowledge, these works do not systematically study the downstream effects mixed human and automatically labeled datasets can have, nor do they propose ways to mitigate such potential issues.

3 Methods

In this section we describe the design and implementation of our case study, from data collection to data annotation and the downstream analysis and briefly present the architecture and workflow of BATS. We additionally present our proposed workflow for mitigating batch effects derived from mixed labels.

3.1 Data acquisition

The dataset designed for our case study was built in house and imaged using a Transmission Electron Microscope (TEM) from liver section of LPP rats, an

established animal model for Wilson's disease [7,9,16,23]. Two populations were acquired, the first consisting of Wild-type LPP rats serving as controls, while the second of *Atp7b* knockout LPP rats at various stages of disease, ranging from still healthy to severely diseased [8]. The LPP rats, originally described in [2], were bred and housed at an institutional animal facility under established laboratory animal care guidelines and in accordance with relevant national legislation and ethical approvals. TEM images were acquired as described in [12]. Our collected dataset includes 277 images deriving from 18 subjects, eight diseased and ten healthy, which results in 8,818 mitochondria present in the dataset.

3.2 Standard AI-assisted workflow: Data curation with BATS

We define the standard AI-assisted workflow as follows: a FM is used to produce initial labels (pre-labels) for a given dataset, which is then manually curated (corrected) by the a human expert. In our study, label generation and data curation was performed using BATS. This tool adopts a client–server architecture in which expert annotators interact with a local napari-based client, while foundation-model inference is performed locally or served remotely [26]. Integrated data-centric mechanisms, such as selective sample confirmation and multi-annotator consensus, prioritize annotation quality. A modular server framework enables extension of segmentation models, ensuring rapid adaptation to new tasks without interface modifications. For the purpose of this work, first, the Cellpose model (cyto) was used for pre-labeling, to generate the initial labels for the data [28]. Next, the expert biologist opened each image with the viewer, visualized the outputs of the model and manually performed any necessary correction to the instance segmentation masks, i.e. delete, add, or correct mitochondria. Additionally, class labels were added for each mitochondria, ranging from one (healthy) to four (severely damaged). Classes one and two can be considered healthy mitochondria while three and four are considered mitochondria damaged by Wilson's disease. After each segmentation was carefully corrected, it, along with its corresponding image, were moved to the curated dataset. BATS includes an additional AI-assisted labeling feature, where bounding boxes or points can prompt the SAM model [13] to refine object labels or generate new ones, but this feature was not used in the current setup. The final dataset consists of a mix of *added instance* labels (when a mitochondria was manually added by the user), *corrected instance* labels (when the mitochondrial mask was manually edited by the user) and *untouched instance* labels (when the Cellpose output was unchanged).

3.3 Proposed AI-assisted workflow: Correcting batch effects derived from mixed labels

In our proposed AI-assisted workflow we introduce an additional step after the standard workflow for data curation. With the masks obtained from Section 3.2, we computed bounding boxes around each mitochondria and used these as prompts to the SAM2 model [25]. In this way a second set of labels was

created for each object, where the instance masks are replaced with the output of the SAM2 model. Therefore, for all images in our dataset two sets of labels are present: those acquired directly from BATS, which includes a mix of model, human corrected and human added labels, which we hereby refer to as **mixed** labels, and those produced using the SAM2 model, which we hereby refer to as **harmonized** labels.

3.4 Downstream task A: Feature extraction & analysis

Using the label sets described in Section 3.3, we were able to extract two sets of features from our data, mixed-derived and harmonized-derived. A total of 17 shape-related features were collected to enable downstream analysis, such as area, perimeter, circularity etc.. We dropped all mitochondria touching the image border, as their shape is not representative. The resulting dataset consists of the following class distribution: class one has 4,361 instances, class two has 71 instances, class three has 1,762 instances and class four has 574 instances.

3.5 Downstream task B: Fine-tuning the segmentation model

A natural downstream task following data curation is to retrain the original foundation model to obtain more accurate segmentations for the given problem. We therefore proceeded to fine-tune the Cellpose 'cyto' model using both labels sets, with a ten-four-four split for train, validation and test set stratified by animal and as much as possible balancing diseased and healthy animals in each set. Both models were trained identically using min-max normalization and resizing to a maximum dimension of 512. Training was performed for 2000 epochs using a learning rate of 1e-5, a weight decay of 0.1, and a batch size of 16, with SGD. The resulting models are referred to as mixed model and harmonized model, reflecting the source of the training labels.

4 Results

In this section we show how mixed labeling affects downstream analysis and segmentation and how our proposed workflow mitigates these effects.

4.1 Batch effect present in mixed label set

In Fig. 2 we clearly see the effect of the standard workflow for AI-assisted labeling on the mixed features. This is most evident in the eccentricity and solidity features, since these are heavily dependent on the smoothness of the segmentation at the boundaries of the object. Small protrusions and concavities occurring around the border can lead to a lower eccentricity (measures the elongation of an object), while over- and under-segmentations can also result in a lower solidity (measures the 'convexity' of an object). The eccentricity-circularity plot (Fig. 2.A) shows three distinct curves attributed to the provenance of instances present

in the data (untouched, corrected, and added). The *untouched instances* display a lower eccentricity and solidity, while *corrected instances* lie at the interval between *untouched* and *added instances*. Cellpose masks often have rougher and cruder boundaries, partially attributed to the upsampling of the model output to the input size (BATS downsamples the image to max. dimension 512 before passing it to the model). *Added instances* typically have smoother boundaries and this curve matches the curve of the harmonized-derived features (circularity at around 0.85 for both curves). In Fig. 2C-D, the curves become uniform and the batch effects disappear, supporting our proposed workflow which smoothens the mixed data effect in AI-assisted workflows. We also analyzed all features stratifying by mitochondria class and metadata and found no other bias attributing to the batch effect presented here. Paired, instance-level analysis revealed that harmonization affects instance types differently, with the largest shifts observed in *untouched* masks (median Δ circularity = 0.192), followed by *corrected* (0.099), while *added instances* remained largely unchanged (0.012). The mean IoU between mixed and harmonized masks was 0.901, indicating that SAM2 primarily refines boundaries rather than altering object identity.

Table 1. Comparing boundary performance on the test set for mixed and harmonized models. Scale defines the distance thresholds used when evaluating how well predicted boundaries match ground-truth boundaries (distance transform computed for both; for more information refer to the Cellpose API [28]).

Label Set	**mixed**			**harmonized**		
Scale(px)	0.5	1	2	0.5	1	2
Precision(%)	90.4 ± 6.6	95.2 ± 4.3	98.3 ± 2.6	$\mathbf{91.0 \pm 6.9}$	$\mathbf{95.3 \pm 5.0}$	98.3 ± 3.0
Recall(%)	82.6 ± 13.8	89.2 ± 10.4	96.4 ± 4.2	$\mathbf{82.9 \pm 11.9}$	$\mathbf{89.6 \pm 8.2}$	$\mathbf{96.8 \pm 3.3}$
F1 score(%)	85.8 ± 9.8	91.8 ± 7.0	97.3 ± 3.1	$\mathbf{86.3 \pm 8.3}$	$\mathbf{92.2 \pm 5.6}$	$\mathbf{97.5 \pm 2.5}$

4.2 Label noise introduced when fine-tuning with mixed labels

In the context of our study, we believe that reporting evaluating metrics of the performance of our models is of little consequence, as the inferred labels are compared against the chosen ground truth and therefore good or bad performance is dependent on the quality of the labels themselves [14]. Indeed, it is the ground truth itself used in standard workflows that we are questioning here, and thus urge the researcher to examine carefully before using for any downstream task. A qualitative evaluation on the other hand, as seen in Fig. 3 where we compare masks from the mixed and harmonized label sets, as well as model outputs on the test set trained on these label sets, indicate that the modeling styles of each label set are picked up by the model during training. Mixed masks, mainly *untouched instances*, have a rougher perimeter, sometimes also going outside the border, compared to the harmonized masks which are smoother, more precisely following the perimeter of the objects. These characteristics are also picked up by the

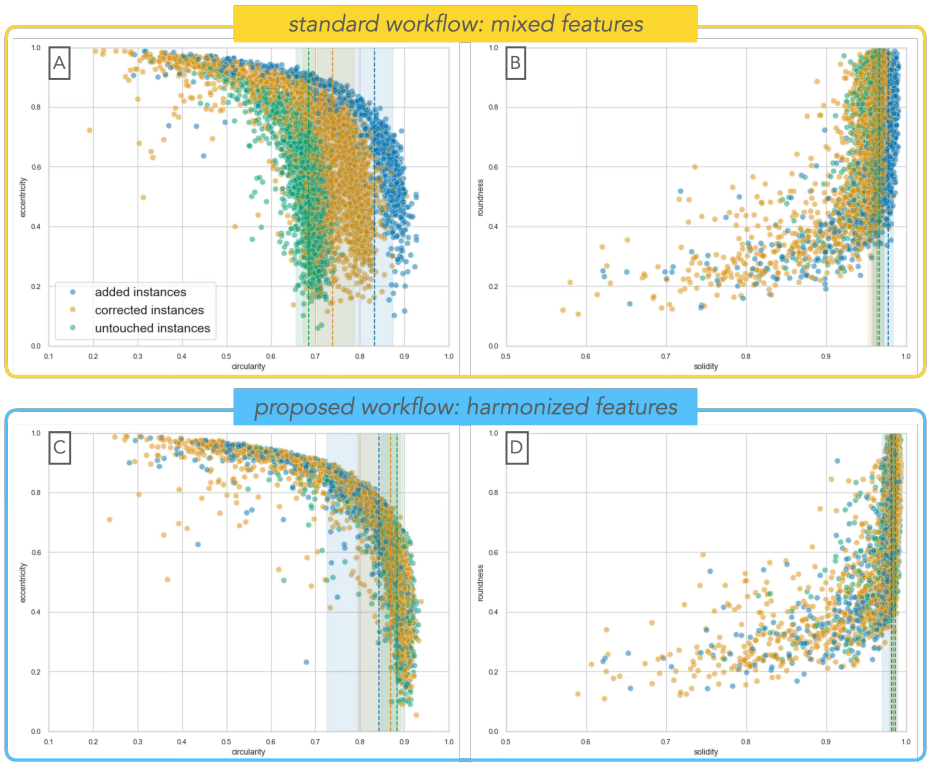

Fig. 2. Scatter plots of features illustrating the batch effect introduced by mixed human- and model-generated labels. On the top panel (standard workflow) scatter plots of features derived from the mixed labels (A-B), in the bottom panel (proposed workflow) scatter plots of features derived from the harmonized labels (C-D). Plots A and C display eccentricity vs. circularity, while B and D display roundness vs. solidity. The batch effect is most apparent in the eccentricity vs. circularity plots; in plot A three distinct curves emerge corresponding to the different instance categories: newly added by the user (blue curve, legend *added instances*), manually corrected model outputs (orange curve, legend *corrected instances*), and unchanged model predictions (green curve, legend *untouched instances*). Notably, the *corrected instances* distribution lies between the *added instances* and *untouched instances* distributions, suggesting an intermediate annotation style resulting from partial human correction of model-generated labels. In the harmonized features the batch effects disappear. Dashed lines and shaded regions indicate medians and interquartile ranges (25th–75th percentiles) for each instance category, highlighting the distribution shifts observed between mixed and harmonized feature representations.

model during fine-tuning, stressing the need to correct batch effects produced from standard AI-assisted workflows. In Tab. 1 we also report metrics on the pixel-level boundaries of the mitochondria, which shows a slight, yet consistent, higher agreement between model and ground truth for the harmonized case, indicating how the mixed model becomes confused when tracing the border due to labeling noise introduced by the mixed labels and results in lower precision, recall and F1-scores.

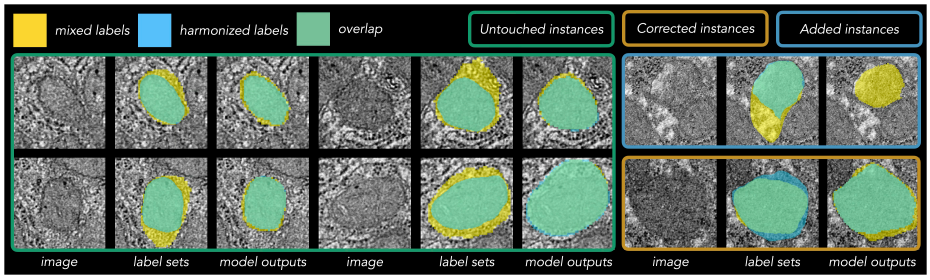

Fig. 3. Different examples of mitochondria and their masks. An overlay of mixed labels (yellow) and harmonized (blue) is shown as well as models outputs on the test set. The bounding boxes around the images indicate the instance type. The *added instance* on the top row (far right) shows an error case of the SAM2 model in which case the harmonized model also misses to segment the object; only 10/8,818 failure cases were identified (morphological changes to the mask, unrelated to the mitochondria boundary and identified after manual inspection) in the data.

5 Discussion

This works demonstrates the issues which can arise from standard AI-assisted labeling workflows, leading to a mix of human and AI-generated labels that can have an impact on downstream tasks (Fig. 1; standard workflow). In our work, we show how shape related features, mainly circularity and solidity, are affected by this mix producing batch effects (Fig. 2), and how downstream analyses based on these features can be affected. Additionally, in some cases the pre-labeling model may fail for a specific type of data (e.g. unseen class, different animal age etc.) requiring human correction only for that specific subset of instances, which would then lead to the discussed batch effects introducing further biases. These effects are of course also dependent on the original model used for segmentation and their presence cannot be attributed to the human annotator during correction, as the visual effects are very slight and difficult to distinguish on an image level, as can be seen in Fig. 3. We describe a strategy to mitigate these effects, using SAM2 to derive masks with box prompts drawn from the mixed labels (Fig. 1; right). Naturally, the effectiveness of this solution depends on the performance of SAM2, which in our case worked almost seamlessly (error=0.11%),

however, we cannot guarantee that it would do so for all biomedical tasks. Nevertheless, we believe that many labs and clinics are rapidly adopting AI-assisted pipelines in their workflow and therefore urge the community to approach the data produced thereof with caution. While we demonstrate the effects of mixed labeling with a case study in microscopy segmentation, there are numerous tasks in the biomedical domain where morphology and border precision play an instrumental role in decision making processes and scientific discovery. To name but one, in tumor delineation for treatment planning, millimeter precision is required when segmenting gliomas in brain MRI, which is used for radiotherapy planning [4,5]. An over-segmentation can lead to the irradiation of healthy tissue, while under-segmenting can leave malignant cells untreated. We therefore aspire that this work will instigate a previously neglected study of datasets produced with AI-assisted workflows, especially in cases where border accuracy is of high importance.

Acknowledgments. We would like to thank Isra Mekki, Gerome Vivar, and Nicholas A. Del Grosso for their assistance and support during development of the BATS software.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adam, G.A., Chang, C.H.K., Haibe-Kains, B., Goldenberg, A.: Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. In: Machine learning for healthcare conference. pp. 710–731. PMLR (2020)
2. Ahmed, S., Deng, J., Borjigin, J.: A new strain of rat for functional analysis of pina. Molecular brain research **137**(1-2), 63–69 (2005)
3. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification; 2021. arXiv preprint arXiv:2101.05224 (2024)
4. De Biase, A., Sijtsema, N.M., Janssen, T., Hurkmans, C., Brouwer, C., van Ooijen, P.: Clinical adoption of deep learning target auto-segmentation for radiation therapy: challenges, clinical risks, and mitigation strategies. BJR| Artificial Intelligence **1**(1), ubae015 (2024)
5. Diana-Albelda, C., García-Martín, Á., Bescos, J.: A review on deep learning methods for glioma segmentation, limitations, and future perspectives. Journal of Imaging **11**(8), 269 (2025)
6. Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., Mehta, P., Li, W., Flores, M., et al.: Monai label: A framework for ai-assisted interactive labeling of 3d medical images. Medical Image Analysis **95**, 103207 (2024)
7. Einer, C., Munk, D.E., Park, E., Akdogan, B., Nagel, J., Lichtmannegger, J., Eberhagen, C., Rieder, T., Vendelbo, M.H., Michalke, B., et al.: Arbm101 (methanobactin sb2) drains excess liver copper via biliary excretion in wilson's disease rats. Gastroenterology **165**(1), 187–200 (2023)

8. Engler, J., Kim, E.J., Kim, D., Shibata, N.M., Lynderup, E.M., Vendelbo, M.H., Akdogan, B., Sailer, J., Fontes, A., Eberhagen, C., et al.: Rapid transcellular hepatic copper depletion by arbm-101 rescues severe liver damage in wilson disease rodents. Biomedicine & Pharmacotherapy **193**, 118867 (2025)
9. Fontes, A., Pierson, H., Bierła, J.B., Eberhagen, C., Kinschel, J., Akdogan, B., Rieder, T., Sailer, J., Reinold, Q., Cielecka-Kuszyk, J., et al.: Copper impairs the intestinal barrier integrity in wilson disease. Metabolism **158**, 155973 (2024)
10. Garcia-Sineriz, I., Giraldez Suarez, J., Galvan Prieto, M., Garcia-Elcano, L., Garcia-Mato, D., Bossa, M., Berto, A., Giraldo Del Viejo, E., Diez-Lopez, C., Lamata, P., et al.: Better labels, better models: Ai-assisted label refining approach improves segmentation performance and annotation efficiency. European Heart Journal-Cardiovascular Imaging **27**(Supplement_1), jeaf367–010 (2026)
11. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain informatics **3**(2), 119–131 (2016)
12. Kabiri, Y., Eberhagen, C., Schmitt, S., Knolle, P.A., Zischka, H.: Isolation and electron microscopic analysis of liver cancer cell mitochondria. In: Mitochondrial Medicine: Volume 3: Manipulating Mitochondria and Disease-Specific Approaches, pp. 277–287. Springer (2021)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. Kofler, F., Wahle, J., Ezhov, I., Wagner, S.J., Al-Maskari, R., Gryska, E., Todorov, M., Bukas, C., Meissen, F., Peng, T., et al.: Approaching peak ground truth. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–6. IEEE (2023)
15. Kumar, S., Datta, S., Singh, V., Datta, D., Singh, S.K., Sharma, R.: Applications, challenges, and future directions of human-in-the-loop learning. IEEE Access **12**, 75735–75760 (2024)
16. Lichtmannegger, J., Leitzinger, C., Wimmer, R., Schmitt, S., Schulz, S., Kabiri, Y., Eberhagen, C., Rieder, T., Janik, D., Neff, F., et al.: Methanobactin reverses acute liver failure in a rat model of wilson disease. The Journal of clinical investigation **126**(7), 2721–2735 (2016)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
18. Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications **9**(1), 5217 (2018)
19. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
20. Northcutt, C., Jiang, L., Chuang, I.: Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research **70**, 1373–1411 (2021)
21. Pandey, R., Purohit, H., Castillo, C., Shalin, V.L.: Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. International Journal of Human-Computer Studies **160**, 102772 (2022)
22. Pavoni, G., Corsini, M., Ponchio, F., Muntoni, A., Edwards, C., Pedersen, N., Sandin, S., Cignoni, P.: Taglab: Ai-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages. Journal of field robotics **39**(3), 246–262 (2022)

23. Polishchuk, E.V., Merolla, A., Lichtmannegger, J., Romano, A., Indrieri, A., Ilyechova, E.Y., Concilli, M., De Cegli, R., Crispino, R., Mariniello, M., et al.: Activation of autophagy, observed in liver tissues from patients with wilson disease and from atp7b-deficient animals, protects hepatocytes from copper-induced apoptosis. Gastroenterology **156**(4), 1173–1189 (2019)
24. Radeta, M., Freitas, R., Rodrigues, C., Zuniga, A., Nguyen, N.T., Flores, H., Nurmi, P.: Man and the machine: Effects of ai-assisted human labeling on interactive annotation of real-time video streams. ACM Transactions on Interactive Intelligent Systems **14**(2), 1–22 (2024)
25. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
26. Sofroniew, N., Lambert, T., Bokota, G., Nunez-Iglesias, J., Sobolewski, P., Sweet, A., Gaifas, L., Evans, K., Burt, A., Doncila Pop, D., et al.: napari: a multi-dimensional image viewer for python. Zenodo (2022)
27. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems **34**(11), 8135–8153 (2022)
28. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. Nature methods **18**(1), 100–106 (2021)
29. Tarsoly, S., Galambos, P., Károlv, A.I.: Comparison of manual and ai-assisted labeling techniques in pixel-wise instance segmentation. In: 2025 IEEE 23rd World Symposium on Applied Machine Intelligence and Informatics (SAMI). pp. 000115–000122. IEEE (2025)