Arctique: An artificial histopathological dataset unifying realism and controllability for uncertainty quantification

Jannik Franzen[1,2,6,*,✉], **Claudia Winklmayr**[1,*], **Vanessa E. Guarino**[1,5,*], **Christoph Karg**[1,*], **Xiaoyan Yu**[1,5], **Nora Koreuber**[1], **Jan P. Albrecht**[1,5], **Philip Bischoff**[2,3,4], **Dagmar Kainmueller**[1,6,✉]

[1] Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association and Helmholtz Imaging, Berlin, Germany
[2] Charité—Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Pathology, Berlin, Germany
[3] Berlin Institute of Health at Charité—Universitätsmedizin Berlin
[4] German Cancer Consortium, German Cancer Research Center, Partner Site Berlin
[5] Humboldt-Universität zu Berlin, Faculty of Mathematics and Natural Sciences, Berlin, Germany
[6] Digital Engineering Faculty of the University of Potsdam

✉ {firstname.lastname}@mdc-berlin.de * equal contribution

Abstract

Uncertainty Quantification (UQ) is crucial for reliable image segmentation. Yet, while the field sees continual development of novel methods, a lack of agreed-upon benchmarks limits their systematic comparison and evaluation: Current UQ methods are typically tested either on overly simplistic toy datasets or on complex real-world datasets that do not allow to discern true uncertainty. To unify both controllability and complexity, we introduce Arctique, a procedurally generated dataset modeled after histopathological colon images. We chose histopathological images for two reasons: 1) their complexity in terms of intricate object structures and highly variable appearance, which yields challenging segmentation problems, and 2) their broad prevalence for medical diagnosis and respective relevance of high-quality UQ. To generate Arctique, we established a Blender-based framework for 3D scene creation with intrinsic noise manipulation. Arctique contains 50,000 rendered images with precise masks as well as noisy label simulations. We show that by independently controlling the uncertainty in both images and labels, we can effectively study the performance of several commonly used UQ methods. Hence, Arctique serves as a critical resource for benchmarking and advancing UQ techniques and other methodologies in complex, multi-object environments, bridging the gap between realism and controllability. All code is publicly available, allowing re-creation and controlled manipulations of our shipped images as well as creation and rendering of new scenes.

1 Introduction

The crucial importance of reliable UQ for the deployment of segmentation algorithms to safety-critical real-world settings has long been recognized by the machine learning community, and the field has seen substantial development of methodology over past years (see e.g. [12, 1, 31]). However, there is a glaring lack of comprehensive evaluation of UQ methods, which makes it difficult to contextualize new methods within the existing paradigms, and renders the choice of suitable UQ

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

methods burdensome for practitioners. One reason for the lack of comparative insight is that often UQ methods are developed from theoretical considerations and tested on hand-crafted toy datasets but fail to provide meaningful, interpretable results on complex real-world datasets [20, 6].

Towards more insightful benchmarking of UQ methods, it is desirable to establish benchmark datasets with ground-truth uncertainty. However, in real-world settings, ground truth uncertainty is usually unattainable. Thus related works have resorted to empirically obtained (and therefore not fully quantifiable and/or controllable) distribution shifts and label noise [20, 3], which has greatly advanced the field, albeit by construction still does not facilitate comprehensive insight into method behavior. Synthetic data generation offers a promising avenue towards improved insight by providing clearly defined data properties and annotations (see [17] for an example from the realm of Explainable AI). However, previous synthetic data generation methodologies proposed in the context of challenging image segmentation problems either excel in controllability but fall short in complexity [30, 39], or vice-versa aim at improved complexity and realism but at the cost of falling short in controllability [8, 41], the latter because learnt image generation, while able to offer some level of conditioning on sought image properties, neither provides full control nor full insight into the image generation process.

To address this gap, we introduce Arctique (ARtificial Colon Tissue Images for Quantitative Uncertainty Evaluation), a procedurally generated histopathological dataset designed to mirror the properties of images derived from H&E stained colonic tissue biopsies, as acquired routinely for safety-critical medical diagnoses in clinical practice [36]. Histopathological images offer a rich and challenging landscape for the application of advanced machine learning methodology, particularly in segmentation [2, 25]. The essential task of accurately delineating and classifying cellular structures is challenging even for trained professionals, due to many sources of uncertainty, e.g. overlapping structures, partial information from the underlying physical tissue-slicing process, and the inherent variability of biological tissues. The demanding nature of this task is reflected in the relative scarcity of fully annotated real-world data sets and high inter-annotator variability (see e.g [14]). Arctique offers the creation of realistic synthetic histopathological images at full controllability, allowing users to manipulate a range of easily interpretable parameters that effectively serve as "sliders" for image- as well as label uncertainties.

Arctique provides 50,000 pre-rendered 512x512-sized images for training and evaluation of segmentation tasks, shipped with exact masks (2D and 3D), metadata storing characteristics of cellular objects, and rendering parameters to re-generate scenes. Furthermore, Arctique provides two main avenues for the controlled study of uncertainty: (1) a blender-based generation framework, which allows to re-generate and manipulate scenes, and (2) a data loader for post-processing images and emulating noisy labels. To assess Arctique's degree of realism, we show that segmentation networks trained exclusively on Arctique can achieve promising zero-shot performance on real H&E images, proving its ability to learn meaningful semantic attributes.

To showcase how Arctique can be used for insightful benchmarking of UQ methods, we assess foreground-background segmentation and semantic segmentation and measure the effect of uncertainty in the images and the labels separately. We benchmark the performance of four widely used UQ-methods (Maximum Softmax Response (MSR), Test Time Augmentation (TTA [38]), Monte-Carlo Dropout (MCD [11]) and Deep Ensembles (DE [22])). For each uncertainty scenario we measure model performance, predictive uncertainty, epistemic uncertainty and aleatoric uncertainty. Overall, we find that our manipulations increase predictive uncertainty in all four benchmarked UQ methods. In particular, we find that their aleatoric uncertainty components mostly track our devised label-level manipulations while their epistemic components mostly track our devised image-level manipulations. This serves as proof-of-concept that Arctique facilitates meaningful and comprehensive UQ benchmarking. Arctique was rendered and assessed on an internal resource of Nvidia A40 GPUs. Our work amounted to an estimated total of 150 GPU-hours.

Dataset access The current version of our dataset, as well as the complete version history, can be accessed via https://doi.org/10.5281/zenodo.11635056. We provide access to 50,000 training and 1,000 test images along with their corresponding instance and semantic masks, including 400 additional exemplary variations corresponding to 50 of the test images. We also provide the dataset used for the experimental results presented in this paper as well as the respective noisy variations. The complete codebase containing scripts for dataset generation, model training and uncertainty estimation is available on GitHub: https://github.com/Kainmueller-Lab/arctique

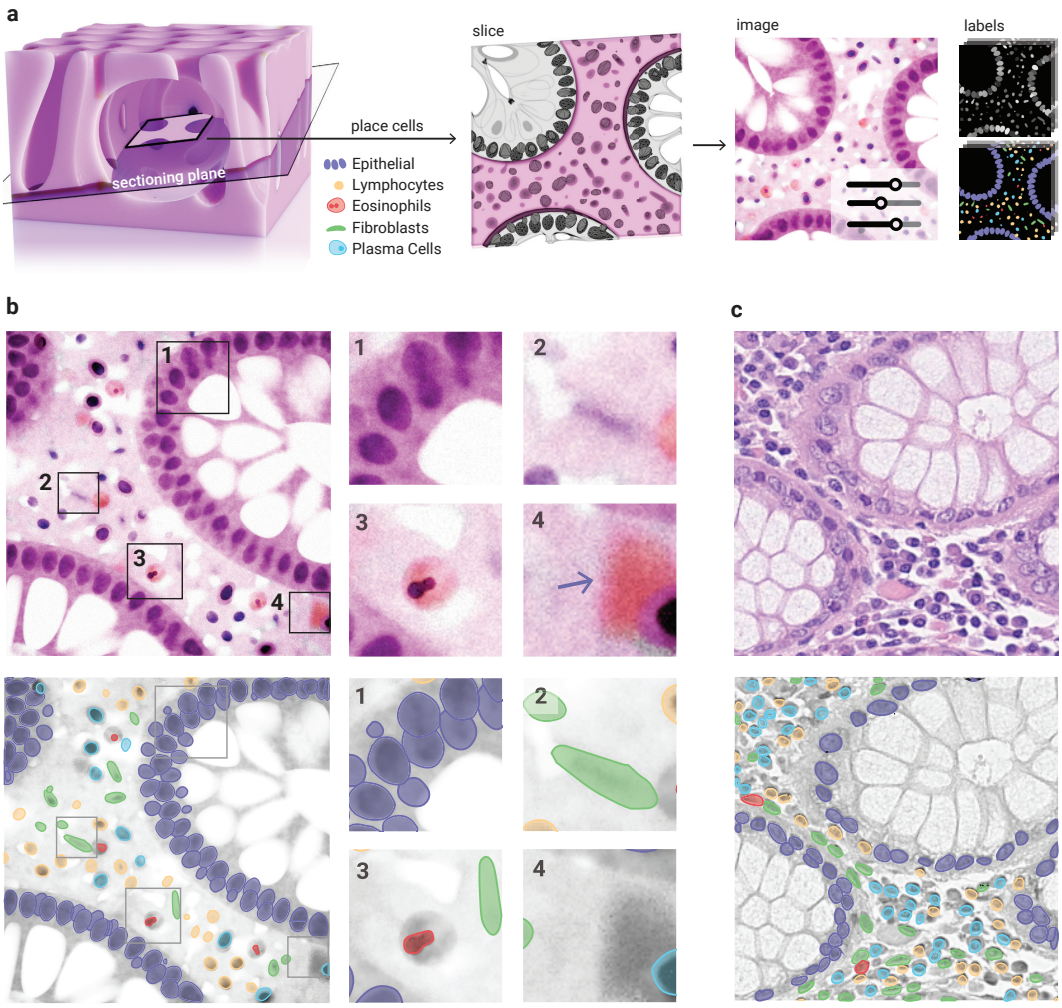

Figure 1: **Generation Process**: (a) To generate complex microscopic images, Arctique artificially replicates the H&E colon image creation protocol. From left to right: Initially, the colonic macrostructure (i.e., the outer epithelial layer) is constructed. This geometry is then artificially sliced, cell nuclei and other objects are placed, and the resulting scene is rendered along with its corresponding 3D stack of instance and semantic masks. (b) The result is a synthetic image (top) with corresponding semantic and instance mask (bottom) featuring numerous cell nuclei that (1) overlap, (2) lie outside the focal plane, (3) exhibit distinct characteristics, and (4) can be confused with perturbing elements. (c) A typical image of a natural H&E stained slice of colonic tissue (top) and the corresponding segmentation (bottom). The *epithelium* exhibits the characteristic flower-like structures called *crypts*. The *stroma* is the densely populated tissue between epithelial crypts.

2 The Dataset

Histopathological Hematoxylin & Eosin (H&E) stained colon tissue images captured under light microscopy pose a significant challenge for segmentation models due to their inherent complexity, manifested by intricate cellular structures, varying staining intensities, and irregular tissue boundaries. The scarcity of exact annotated data further aggravates the problem, hindering the development of robust segmentation models. Our synthetic dataset Arctique mimics the complexity of real H&E stained colon tissue images akin to those in the *Lizard* dataset [14], capturing the diverse cellular morphology, nuanced staining patterns, and irregular tissue boundaries present in real histopathological samples. To this end, we devised a Python-based image generation pipeline on top of the 3D ray-tracing software Blender. This approach, as opposed to the alternative of relying

on generative models, ensures controllability (and reproducibility) while allowing for the creation of realistic scenes. Consequently, each of our 50,000 pre-rendered images and labels can be easily recreated and subjected to controlled modifications. We generate each H&E image along with its corresponding masks via the following procedural data generation pipeline, as shown in Figure 1a:

1. Macro-structure: We start with generating a 3D model representing the characteristic topology of the colon tissue architecture. Specifically, we focus on the *epithelial crypts*, which do not only follow an intriguing pattern (see Figure 1a left) but are also of pathological relevance. For example, colon cancer typically originates at this outer layer of the colon. We create the outer tissue topology by modeling the rod-shaped crypts in a densely packed hexagonal pattern, as depicted in Figure 1a (for details see Appendix).

2. Placing of cells: Next, we populate our scene with five predominant cell types. Within the stroma, the connective tissue between the crypts, we randomly distribute plasma cells, lymphocytes, eosinophils, and fibroblasts. The cells of the epithelial crypts, specifically epithelial cells and corresponding goblet cells (white bubbles, see Figure 1), are placed according to a 3D adaptation of the Voronoi cell generation algorithm (cf. [37, 23] and Appendix). Each cell model includes both its nucleus and the surrounding cytoplasm. For instance, Figure 1b3 illustrates the peanut-shaped nucleus of an eosinophil within its reddish-stained cytoplasm. Each cell type is characterized by controllable parameters such as typical diameter, elongation, and nucleus shape. A comprehensive description of the parameter sampling methodology is provided in the Appendix.

3. Sectioning: A significant source of complexity arises from the fact that histopathological images are 2D slices of the underlying 3D architecture. To replicate this, we digitally slice through our 3D macro structure and cells. This approach ensures that the visible features of cells vary depending on their location and orientation relative to the sectioning plane. For example, in Figure 1b2, we can faintly observe two fibroblasts: one cut along its major axis and another cut vertically.

4. Staining: In real-world histopathological images, the staining colors are derived from Hematoxylin & Eosin (H&E) staining. Hematoxylin binds to DNA in the nucleus, giving it a purple appearance, while eosin stains the surrounding tissue architecture in a reddish-purple hue (see Figure 1c). To replicate this, we model digital staining of the cell cytoplasm, cell nuclei, and surrounding tissue using controllable parameters such as staining hue, staining intensity, and inherent staining intensity noise. This is achieved using Blender-specific shaders, as detailed in the Appendix.

5. Rendering: The final scene is rendered using ray tracing from a virtual camera positioned above the light source and tissue slice (see Appendix). By adjusting the camera's focal plane, we achieve a depth-blurring effect characteristic of histopathological light microscopy images. This workflow enables the generation of both 2D images and high-resolution 3D stacks. Moreover, synthetic image generation provides precise pixel-wise semantic- and instance masks, serving as exact ground truth.

Parameters: Various parameters can be gradually adjusted to control the rendering process and allow for precise customization of the generated images: *Cell shapes*: Adjustments include cell diameter, elongation, bending, and shape noise for linear interpolation between cell types. *Cell distribution*: Parameters cover cell locations, occurrence ratios of cell types, and cell density in the stroma. *Tissue parameters*: Configurations include tissue thickness and degrees of tearing. *Staining parameters*: Settings include staining colors and intensities for cells, nuclei, and tissue. By conducting statistical analysis of the Lizard dataset [14] and consulting with a pathologist, we fine-tune these parameters to closely align with the images from the Lizard dataset.

Assessment of Realism: We assess the resemblance between Arctique and real data through two complementary approaches: a qualitative analysis of the detail captured by Arctique, and a quantitative evaluation of its effectiveness in training an H&E-based nuclei segmentation and classification model [4, 33], which is then applied zero-shot to real data. Figure 1b demonstrates the fidelity of our dataset in capturing the characteristic nuances observed in real histopathological images, such as those in the Lizard dataset. Figure 1b1 depicts a synthetic ring of densely packed and overlapping epithelial cell nuclei. Our rendering pipeline generates exact corresponding instance masks, and distinguishes the depth of each nucleus within the tissue based on precise depth information. In Figure 1b2, an artificial fibroblast is depicted, exhibiting minimal distinguishability. The pronounced blur is attributed to the cell's deep location within the tissue, a phenomenon commonly observed in real images. Figure 1b3 showcases an artificial eosinophil cell, characterized by its distinctive peanut-shaped nucleus. Our dataset accurately models this characteristic feature, including the reddish hue staining of the

surrounding cytoplasm, which contrasts with other cytoplasmic staining patterns. The resemblance of eosinophil cytoplasm to blood cells, which are also modeled in our dataset (Figure 1b4), presents challenges for pathologists and segmentation models alike.

Zero-shot segmentation of real data: We assess the suitability of Arctique as surrogate training data as follows: We use Arctique to train a HoVer-NeXt (HN) model [4], an architecture that has been shown to yield state-of-the-art results when trained on real H&E data; We then conduct zero-shot inference on real H&E data [14]. To validate Arctique's ability to infer semantically meaningful intermediate attributes, we compare baseline results with an enhanced Arctique version containing randomly injected anomalies, as well as with a simplified version with only pixel depth masks on background. As shown in Figure 2, both F1 score and Hausdorff distance metric (weighted for true positives per class) support Arctique's realism and value. Considering the metrics per class, the synthetic data achieves a positive correlation between predictions and observations for epithelial cells without fine-tuning, while higher significance is observed for lymphocytes when anomalies are used as training data. Qualitative tile inspections further reveal that Arctique can detect previously discordant nuclei. For further details on the training process, datasets description, and metrics comparison, see the Appendix.

3 Benchmarking uncertainty quantification methods

To showcase the capabilities of the Arctique dataset for benchmarking uncertainty quantification methods, we study the effect of image-level and label-level uncertainties on foreground-background segmentation (FG-BG-Seg) and semantic segmentation (Sem-Seg) [24]. To serve as a proof-of-concept, we evaluate the performance of established algorithms on our data, namely segmentation with a UNet backbone [32, 18], and uncertainty estimation with four popular methods, two ensemble-based, namely *Monte-Carlo Dropout* (MCD, [11]) and *Deep Ensembles* (DE, [22]), one heuristic, namely *Test Time Augmentation* (TTA, [38]), and, for comparative purposes, one deterministic model known as *Maximum Softmax Response* (MSR).[1] (See Appendix for implementation details).

In accordance with [20], we use the predictive entropy $\mathbb{H}[Y|x, \mathcal{D}]$, conditional on the training set $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$, as the uncertainty measure of our predictive distribution $p(y|x, \mathcal{D})$, called *predictive uncertainty (PU)*. For all UQ models except MSR, we compute $\mathbb{H}[Y|x, \mathcal{D}]$ in (1) by sampling from the models and averaging over the softmax outputs. Following [7, 21], we can then perform an information-theoretic decomposition to disentangle, respectively, the epistemic and the aleatoric components. In this setting, the epistemic uncertainty is defined as the mutual information between output y and model parameters ω:

$$\underbrace{\mathbb{H}[Y|x, \mathcal{D}]}_{\text{Predictive Unc. (PU)}} = \underbrace{\mathbb{I}[Y; \omega|x, \mathcal{D}]}_{\text{Epistemic Unc. (EU)}} + \underbrace{\mathrm{E}_{p(\omega|\mathcal{D})}[\mathbb{H}[Y|x, \omega]]}_{\text{Aleatoric Unc. (AU)}}. \tag{1}$$

Eq. (1) shows that the aleatoric component correlates with the ambiguity inherent to the data and we should expect high values when there is a mismatch between image and label [21]. In particular, this implies that the aleatoric component is only meaningful for in-distribution data. The epistemic component, on the other hand, correlates with the model's lack of knowledge. It is sensitive to out-of-distribution (OOD) data and can be compensated for by the addition of new training data.

While the UQ measures discussed so far yield pixel-wise results, we want to relate the uncertainty measures to our image- and label-level manipulations and are thus interested to aggregate pixel-level results to obtain uncertainty scores per image. It has been shown in [20] that the specific type of aggregation employed can hugely influence the behavior of uncertainty metrics. To account for this we tested the three aggregation strategies discussed in [20]: *image-level aggregation*, where uncertainty scores for all pixels are summed for each image and averaged over the dataset; *patch-level aggregation*, where uncertainty scores are aggregated within a sliding window and the maximum of the patch scores is taken as the image-level score; and *threshold-level aggregation*, which considers only uncertainty scores above an empirical quantile $\widehat{Q}_{u(\widehat{y})}(p)$ for a chosen uncertainty measure u, then calculates their mean. All results presented in the main text are generated using *threshold-level*

[1] Where a deterministic model predicts a distribution $p(y|x, \omega)$, we define $1-$MRS as $1 - \max_c p(y = c|x, \omega)$, a metric employed as computationally cheap alternative to the predictive entropy [16] despite depending only on a single model realization (see also [27]).

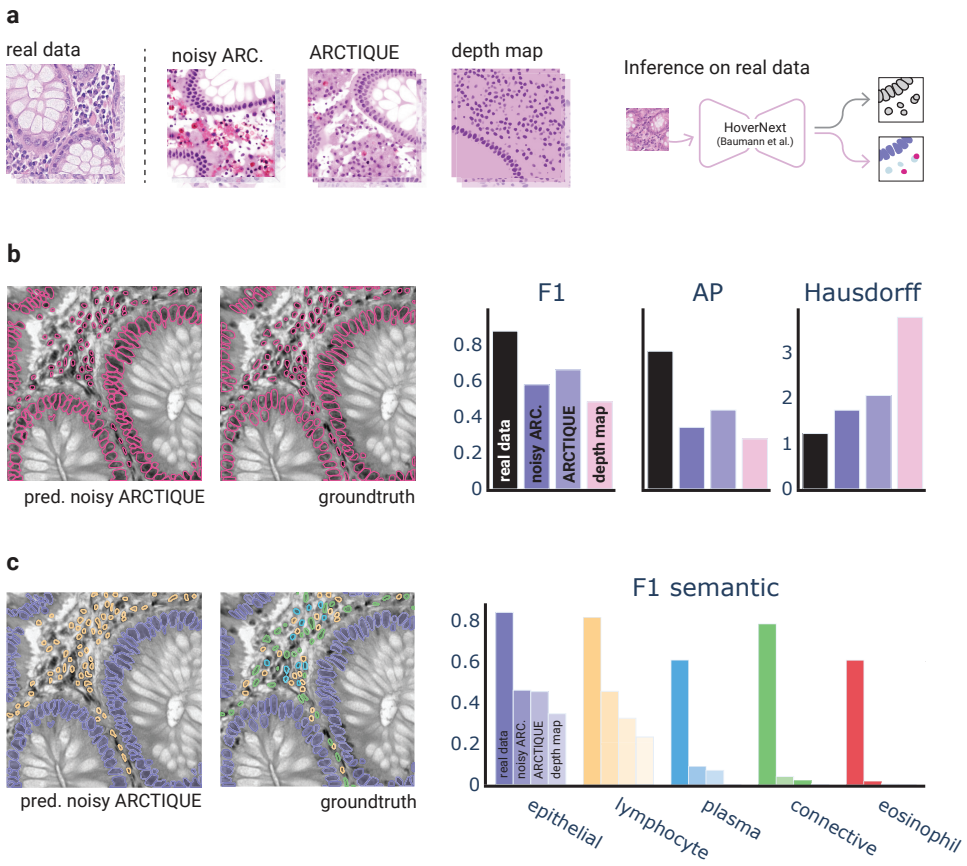

Figure 2: **Inference on the Lizard dataset using HoVer-NeXt (HN) models trained on Arctique:** (a) Graphical illustration of the Arctique variants used for zero-shot learning, arranged on the left by complexity level (from most to least complex and noisy). Each variant aims to enhance the model's generalization across diverse structural and textural details. On the right, a schematic representation depicts the post-processed raw class- and instance map outputs from the HN model during inference. (b) and (c) show visual and quantitative results for instance- and semantic segmentation, respectively, with bar plots comparing the baseline HN model trained on Lizard data (black) to the three HN models trained on simulated datasets of varying complexity. All metrics and predictions are averaged across 5 inference rounds, each with 16 Test-Time Augmentations. Note that the colors of the bars in (c) correspond to the colors of celltypes in the example.

aggregation and normalization based on the image size. In the Appendix, we provide results from alternative aggregation strategies for comparison.

In our experiments, we validate uncertainty measures using the variables defined in Eq. (1). Our approach differs from previous studies, such as [20, 29, 10], by focusing on how well information-theoretic definitions of aleatoric and epistemic uncertainty capture true uncertainty within our dataset, especially for complex tasks like semantic segmentation. While some studies, like [15, 26], examine epistemic uncertainty for segmentation, they are generally limited to in-distribution data. Additionally, we track prediction accuracy across varying noise levels, expecting accuracy to decrease as overall uncertainty increases.

3.1 Label Noise

In the first step, we look at the effect of uncertain labels. In biomedical data, this is a common issue as complex and ambiguous images yield high disagreement even among expert annotators. In

real-world images, we should expect some correlation between uncertainty in images and uncertainty in the labels. For example, cells with lower contrast staining might be harder to identify for human annotators leading to more missing labels. However, we believe, that there is a benefit to studying label-noise in isolation as it can give us valuable insights into model calibration and the sensitivity of UQ methods [6, 13, 19].

We devise two types of label-noise tailored to different segmentation tasks: *Sem-Seg:* Class labels are randomly switched. The noise level reflects the probability for each single-cell label to be switched to another class. *FG-BG-Seg:* Labels of single cells are manipulated by shifting, scaling, elastic transform or completely removed (missing label). The noise-level reflects the probability that any singe cell is affected by any of these modifications. Both types of label noise are illustrated in the top row of Figure 3.

Figure 3 summarizes the results of uncertainty evaluation in the presence of label noise: for both segmentation tasks, we find that performance decreases with increasing label noise. This is to be expected as unreliable labels make it harder for the model to learn generalizable patterns. Predictive uncertainty (PU) increases as a function of label noise across both tasks. This confirms that what we would intuitively consider as "making the segmentation task harder" will actually decrease performance and increase uncertainty. Further, we find that across both segmentation tasks the majority of uncertainty stems from the aleatoric component. This is in keeping with the theoretical claim that aleatoric uncertainty mainly captures data-inherent uncertainty.

While the aggregate measures provide a convenient way to assess how training with noisy labels can affect a model's uncertainty, Arctique also comes with exact labels and thus allows to study the impacts of label manipulation on the pixel-level. Figure 3 provides a detailed example of what models trained with and without label-noise learn about the training images. In particular, we see that when some cell labels are consistently missing the model will not learn to identify these corresponding cells. However, despite this, uncertainty maps still indicate high uncertainty in regions with missing labels. This observation suggests that uncertainty quantification may offer a solution to address the common issue of sparse annotations in biomedical images. Conversely, in densely packed regions, the model tends to interpolate across missing labels, as seen in the epithelial crypt in the bottom left of 3(c). Here, however, the incorrect merging of cells is not captured in the uncertainty maps, highlighting an area for improvement in uncertainty quantification methods. This duality underscores the value of evaluating UQ methods for their ability to handle both sparse and ambiguous label regions effectively.

3.2 Image Noise

The greatest advantage of having full control over the dataset creation is that it allows us to perform targeted manipulations to certain aspects of the image while leaving all others unchanged. We are thus able to create samples that gradually transform from near-OOD, where outlier and inlier classes are quite similar, to far-OOD, where an outlier is more distinct. This method of generating data is fundamentally different from other common strategies, such as applying augmentations like color shifts or blur [5, 32], where we achieve global manipulations which do not correlate with input features; or testing on images from a different domain [35] (e.g. data from a different organ) where the exact impact on image components is unknown and uncontrolled.

For the manipulation named *Nuclei-Intensity*, we progressively remove staining from the cells' nuclei until they are virtually indistinguishable from their surroundings. This localized process preserves contrast and details in other regions of the image and mimics potential real staining inconsistencies. In contrast, an image-level contrast reduction operation evenly reduces it across the entire image, affecting all elements equally. For the *Blood-Stain* manipulation, we progressively increase both the red stains and the number of blood cells, simulating realistic and extreme variations in blood cell abundance. This adjustment reflects a possible scenario in histology where red-stained artifacts may be mistaken for cell types like eosinophils, which naturally exhibit red staining.

Figure 4 shows examples for both types of image-level manipulations and their effects on model performance and uncertainty. Subfigure 4 (a) shows the impact of manipulating the nuclei-intensity on the FG-BG-Seg task. As might be expected, we observe a decrease in accuracy and an increase in the uncertainty measures as staining intensity decreases. While the aggregated effects may appear subtle, the error maps reveal a clearly visible decline in prediction performance: as staining weakens, the model starts to hallucinate cells in the tissue of the crypt-structures.

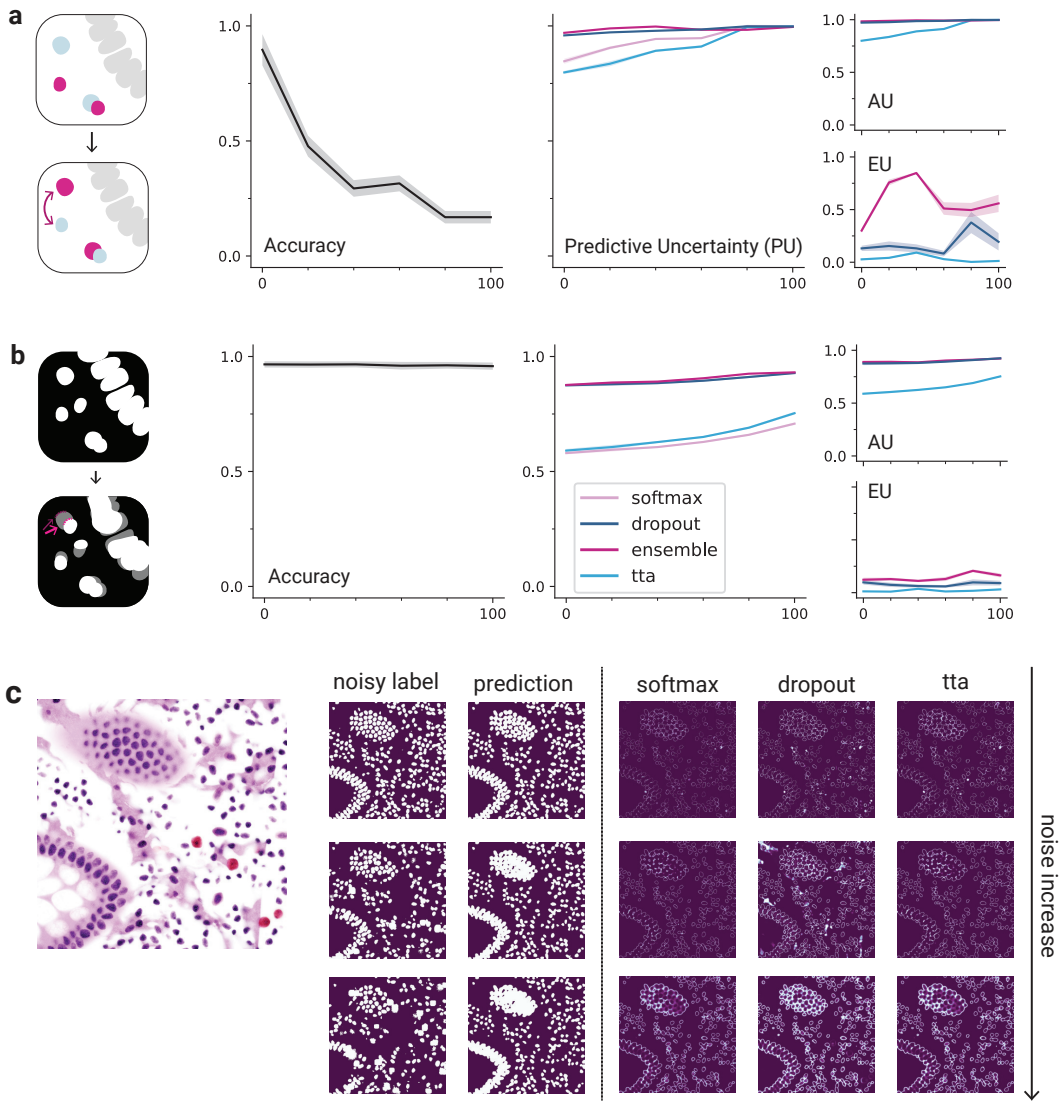

Figure 3: **Illustration of two types of label uncertainty and their effect on model performance and uncertainty measure.** (a) Effect of noisy class labels on Sem-Seg: illustrations on the left show an example of possible label confusion. The two large panels in the middle show model performance across noise levels (x-axis) as measured by accuracy and predictive uncertainty for all four UQ methods. The two smaller panels on the right show aleatoric and epistemic uncertainty for DE, TTA and MCD. (Note that MSR does not permit decomposition, therefore not shown.) (b) Effect of noisy label shapes on FG-BG-Seg: subpanels analogous to (a). (c) Qualitative example of the impact of noisy labels for FG-BG-Seg on prediction performance and how this is captured in the PU maps.

In Subfigure 4 (b) we illustrate the effect of manipulating the the blood-cells and -stains on the Sem-Seg task. During training, the model has learned to identify eosinophil cells by their characteristic red-brick color after staining. As the abundance of blood-cells increases, the model begins to misclassify them as eosinophil cells as is reflected by the qualitative error maps in the top section of 4 (b) and by the aggregated accuracy results in the bottom part of 4 (b), both of which indicate a significant decrease in accuracy. Notably, this decrease in performance is not captured by the uncertainty measures. In fact, the error maps reveal that regions affected by blood cells exhibit particularly low uncertainty values, indicating that high error rates do not correlate with high uncertainty. For DE we

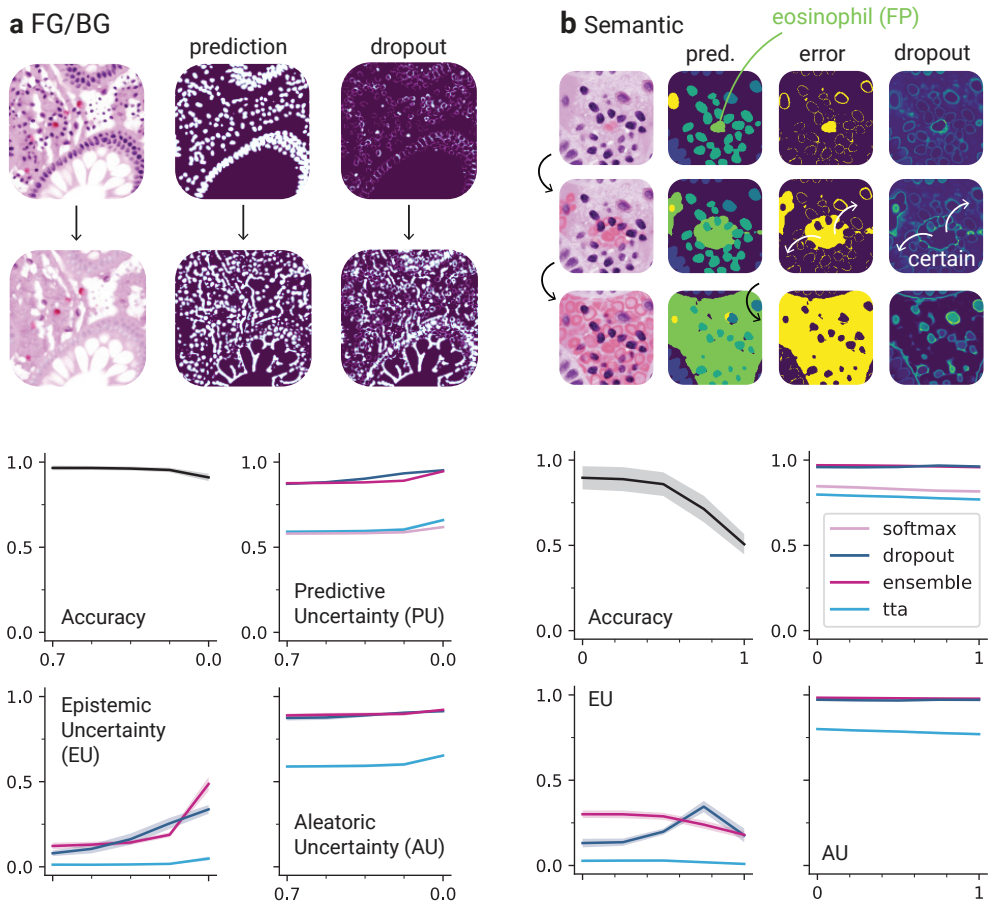

Figure 4: **Illustration of Image-level noise:** (a) Illustration of an image undergoing decreasing intensity of nuclei staining. The small image patches on the top illustrate qualitatively how FG-BG prediction performance and PU (for the example of MCD) are affected as staining is removed. The four panels on the bottom summarize for all four uncertainty methods how accuracy, PU, AU and EU react to the gradual change in staining. (b) illustrates the effect of the increasing prevalence of blood-cells. Similar as in (a) the small image patches on the top show the qualitative changes in semantic prediction performance and uncertainty. Here we additionally show the error maps next to the PU maps to highlight how blood cells are incorrectly identified as eosinophil cells, however the model remains confident in its prediction. The four panels on the bottom are arranged analogous to (a) and further illustrate the decrease in performance while uncertainty remains relatively unchanged.

even observe a slight decrease in the uncertainty as the prevalence of blood cells increases, further emphasizing the weak correlation between error rates and uncertainty values.

We conclude that both experiments demonstrate Arctique's capabilities for in-depth analysis of uncertainty behavior. The two manipulations highlight common challenges encountered in real-world H&E images: without perfect labels, it becomes nearly impossible to ascertain whether high uncertainty values indicate subtle features in the data or stem from a miscalibrated uncertainty model. In the case of the *Nuclei-Intensity* alteration, the uncertainty method effectively identifies a genuine issue. In contrast, with the *Blood-Stain* manipulation, the uncertainty quantification (UQ) models demonstrate inadequacies in correctly calibrating the model.

4 Discussion

UQ carries the promise to increase the reliability of machine learning models so that these models can be more widely deployed even in safety-critical domains. To this end, we must be confident that the UQ methods we develop follow through on their claims. We believe that Arctique constitutes a valuable first step for a more thorough and interpretable evaluation of UQ metrics.

While the domain of histopathology may represent a specialized domain it serves as a valuable testbed due to its versatility and the presence of common uncertainty sources, such as missing or incorrect labels and overlapping instances. Moreover, this domain is particularly relevant for UQ, as it is critical for medical diagnosis and often suffers from incomplete and inaccurate annotations.

Our main goal in this publication is to introduce the Arctique dataset and illustrate its utility for evaluating UQ methods, yet it also opens numerous promising avenues for further research. One important follow-up would be to expand the range of studied UQ methods, particularly in Active Learning (AL), where uncertainty plays a central role in sampling strategies. Recent studies suggest that prioritizing high epistemic uncertainty can improve AL performance, while focusing on aleatoric uncertainty may be less effective ([6], [28]). Arctique's controlled uncertainty levels make it suitable for evaluating AL sampling, integrating uncertainty into optimization, and exploring domain adaptation strategies ([9], [40], [34]). In particular, Arctique allows to straightforwardly combine multiple sources of uncertainty at any level, thus constituting a unique testbed for methodology that seeks to disentangle AU and EU.

Our dataset can also be applied in the context of Explainable AI (XAI) evaluations, where transparent decision-making is crucial for trustworthiness. In contrast to simpler datasets like FunnyBirds [17], which focus on single-class tasks, Arctique offers a realistic multi-object environment. This complexity allows XAI methods to be benchmarked on relevant concepts that reflect the characteristics of real data, and to analyze interpretability and predictiveness across complex co-occurrence patterns, as for example cell nuclei and cytoplasm, Finally, future research could extend our framework with image- and label modifications, encompassing imaging modalities, tissue types, and staining variations. We encourage users to devise their own modifications suitable to their specific evaluation needs. A direct next step could be studying uncertainty for related tasks such as panoptic segmentation or 3D models.

To conclude, our work contributes Arctique, a complex, realistic yet fully controllable dataset of synthetic images, together with a broad range of "sliders" for targeted manipulation. As a proof-of-concept, we show that we can tailor label- and image manipulations such that they are selectively picked up by the aleatoric and epistemic components of established UQ methods, which suggests that Arctique is a valuable resource for UQ methods development and benchmarking, with clear potential for extensions into orthogonal methodological realms like XAI.

Acknowledgments

We wish to thank Aasa Feragen, Kilian Zepf, Paul Jäger and Lorenz Rumberger for inspiring discussions. Funding: German Research Foundation (DFG) Research Training Group CompCancer (RTG2424), DFG Research Unit DeSBi (KI-FOR 5363, project no. 459422098), DFG Collaborative Research Center FONDA (SFB 1404, project no. 414984028), DFG Individual Research Grant UMDISTO (project no. 498181230), Synergy Unit of the Helmholtz Foundation Model Initiative, Helmholtz Einstein International Berlin Research School In Data Science (HEIBRiDS).

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion, 76:243–297, 2021.

[2] Mohammed M Abdelsamea, Usama Zidan, Zakaria Senousy, Mohamed Medhat Gaber, Emad Rakha, and Mohammad Ilyas. A survey on artificial intelligence in histopathology image analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(6):e1474, 2022.

[3] Neil Band, Tim GJ Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. arXiv preprint arXiv:2211.12717, 2022.

[4] Elias Baumann, Bastian Dislich, Josef Lorenz Rumberger, Iris D Nagtegaal, Maria Rodriguez Martinez, and Inti Zlobec. Hover-next: A fast nuclei segmentation and classification pipeline for next generation histopathology. In Medical Imaging with Deep Learning, 2024.

[5] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 9901 of LNCS, pages 424–432. Springer, Oct 2016.

[6] Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27, pages 715–726. Springer, 2021.

[7] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In International conference on machine learning, pages 1184–1193. PMLR, 2018.

[8] Kexin Ding, Mu Zhou, He Wang, Olivier Gevaert, Dimitris Metaxas, and Shaoting Zhang. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. Scientific Data, 10(1):231, 2023.

[9] Kuan Fang, Yunfei Bai, Stefan Hinterstoisser, Silvio Savarese, and Mrinal Kalakrishnan. Multi-task domain adaptation for deep learning of instance grasping from simulation, 2018.

[10] Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, and Milad Alizadeh. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis. 2019.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning, pages 1050–1059. PMLR, 2016.

[12] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. arXiv preprint arXiv:2107.03342, 2021.

[13] Camila Gonzalez, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn, and Anirban Mukhopadhyay. Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, pages 304–314, Cham, 2021. Springer International Publishing.

[14] Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 684–693, 2021.

[15] Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2020.

[16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. CoRR, abs/1610.02136, 2016.

[17] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 3981–3991, October 2023.

[18] Pavel Iakubovskii. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019.

[19] Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in Neuroscience, 14, 2020.

[20] Kim-Celine Kahl, Carsten T Lüth, Maximilian Zenk, Klaus Maier-Hein, and Paul F Jaeger. Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. arXiv preprint arXiv:2401.08501, 2024.

[21] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems, 30, 2017.

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30, 2017.

[23] Hugo Ledoux. Computing the 3d voronoi diagram robustly: An easy explanation. In 4th International Symposium on Voronoi Diagrams in Science and Engineering (ISVD 2007), pages 117–129, 2007.

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[25] Amirreza Mahbod, Christine Polak, Katharina Feldmann, Rumsha Khan, Katharina Gelles, Georg Dorffner, Ramona Woitek, Sepideh Hatamikia, and Isabella Ellinger. Nuinsseg: a fully annotated dataset for nuclei instance segmentation in h&e-stained histological images. arXiv preprint arXiv:2308.01760, 2023.

[26] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. CoRR, abs/1811.12709, 2018.

[27] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 24384–24394, June 2023.

[28] Vu-Linh Nguyen, Mohammad Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. Machine Learning, 111, 01 2022.

[29] Janis Postels, Mattia Segù, Tao Sun, Luca Daniel Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 17870–17909. PMLR, 17–23 Jul 2022.

[30] Satwik Rajaram, Benjamin Pavie, Nicholas E F Hac, Steven J Altschuler, and Lani F Wu. Simucell: a flexible framework for creating synthetic microscopy images. Nature Methods, 9:634–635, 2012.

[31] Marianne Rakic, Hallee E Wong, Jose Javier Gonzalez Ortiz, Beth Cimini, John Guttag, and Adrian V Dalca. Tyche: Stochastic in-context learning for medical image segmentation. arXiv preprint arXiv:2401.13650, 2024.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

[33] Josef Lorenz Rumberger, Elias Baumann, Peter Hirsch, Andrew Janowczyk, Inti Zlobec, and Dagmar Kainmueller. Panoptic segmentation with highly imbalanced semantic labels. In 2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC), pages 1–4, 2022.

[34] Shaheer U. Saeed, João Ramalhinho, Mark Pinnock, Ziyi Shen, Yunguan Fu, Nina Montaña-Brown, Ester Bonmati, Dean C. Barratt, Stephen P. Pereira, Brian Davidson, Matthew J. Clarkson, and Yipeng Hu. Active learning using adaptable task-based prioritisation, 2022.

[35] Jeppe Thagaard, Søren Hauberg, Bert van der Vegt, Thomas Ebstrup, Johan D. Hansen, and Anders B. Dahl. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I, page 824–833, Berlin, Heidelberg, 2020. Springer-Verlag.

[36] M Titford. The long history of hematoxylin. Biotechnic & histochemistry, 80(2):73–78, 2005.

[37] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy. Scipy 1.0-fundamental algorithms for scientific computing in python. CoRR, abs/1907.10121, 2019.

[38] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing, 338:34–45, 2019.

[39] Veit Wiesmann, Matthias Bergler, Ralf Palmisano, Martin Prinzen, Daniela Franz, and Thomas Wittenberg. Using simulated fluorescence cell micrographs for the evaluation of cell image segmentation algorithms. BMC Bioinformatics, 18(1):176, December 2017.

[40] Mark Woodward and Chelsea Finn. Active one-shot learning, 2017.

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 3836–3847, October 2023.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We mentioned further extensions in the discussion
   (c) Did you discuss any potential negative societal impacts of your work? [No] Because to our knowledge our contribution is unlikely to have negative societal impacts
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The dataset is published on Zenodo while the codebase is available on GitHub. URLs to the repositories are provided at the end of the introduction
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All training details are included in part A.3 of the supplementary material
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We make this statement at the end of the introduction
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We tested our models on the Lizard dataset and cited the original publication
   (b) Did you mention the license of the assets? [Yes] The data and codebase licenses are provided in part D of the supplementary material
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The dataset is published on Zenodo while the codebase is available on GitHub. URLs to the repositories are provided at the end of the introduction
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]