FISHing in Uncertainty: Synthetic Contrastive Learning for Genetic Aberration Detection

Simon Gutwein[1,2,3][0009−0004−8406−0736], Martin Kampel[2][0000−0002−5217−2854], Sabine Taschner-Mandl[1][0000−0002−1439−5301], and Roxane Licandro[3][0000−0001−9066−4473]

[1] St. Anna Children's Cancer Research Institute, Vienna, Austria
[2] TU Wien, Institute of Visual Computing and Human-Centered Technology, Computer Vision Lab, Vienna, Austria
[3] Medical University of Vienna, Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research, ELIA Group, Vienna, Austria

Abstract. Detecting genetic aberrations is crucial in cancer diagnosis, typically through fluorescence *in situ* hybridization (FISH). However, existing FISH image classification methods face challenges due to signal variability, the need for costly manual annotations and fail to adequately address the intrinsic uncertainty. We introduce a novel approach that leverages synthetic images to eliminate the requirement for manual annotations and utilizes a joint contrastive and classification objective for training to account for inter-class variation effectively. We demonstrate the superior generalization capabilities and uncertainty calibration of our method, which is trained on synthetic data, by testing it on a manually annotated dataset of real-world FISH images. Our model offers superior calibration in terms of classification accuracy and uncertainty quantification with a classification accuracy of 96.7% among the 50% most certain cases. The presented end-to-end method reduces the demands on personnel and time and improves the diagnostic workflow due to its accuracy and adaptability. All code and data is publicly accessible at: https://github.com/SimonBon/FISHing.

Keywords: FISH imaging · contrastive learning · gene aberration · uncertainty estimation · aleatoric uncertainty · synthetic data

1 Introduction

Evaluating genetic aberrations is crucial in cancer diagnostics, including breast [22], lung cancer [25] or neuroblastoma [2]. Fluorescence *in situ* hybridization (FISH) followed by fluorescence microscopy, allows for the visualization of gene copies in cell nuclei. Fluorophore-labeled probes are employed to bind to DNA sequences of a gene of interest (*target*) and a *reference* gene and are imaged using distinct wavelengths. FISH signals manifest as red and green spots in an RGB color scheme, with the nucleus stained blue as illustrated in Fig. 1A for *HER2* in breast cancer and Fig. 1B for *MYCN* in neuroblastoma risk stratification [7].

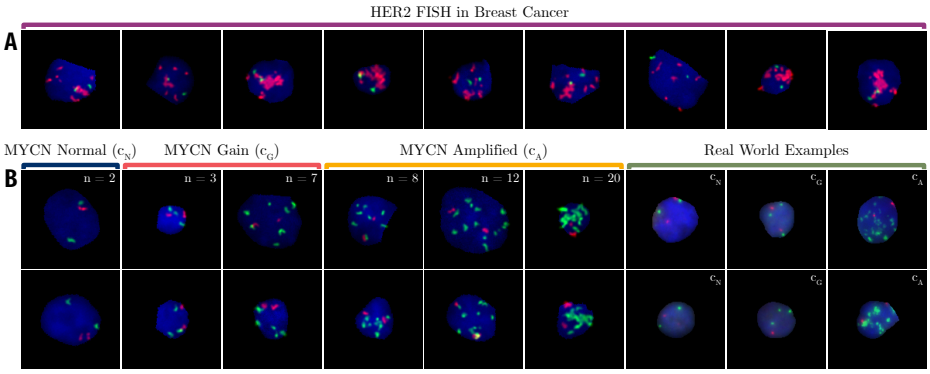

Fig. 1: FISH-image patches with stained cell nuclei in blue image channel for row **A** and **B**. **A**: *HER2* (*target*) as red signals, *CEN17q* (*reference*) as green signals. **B**: Synthetic images of *MYCN* FISH, *MYCN* (*target*) as green signals, *NMI* (*reference*) as red signals, showing diagnostic classes: *MYCN Normal* (c_N), *Gain* (c_G), and *Amplified* (c_A) along real world image examples. n indicates the number of *MYCN* signals.

Challenges: Assessing gene copy number in FISH images requires expert manual evaluation to count the signals. It is a tedious and subjective process, which embodies inherent uncertainty, particularly when signal proximity or cluster formation complicates the precision of gene copy number determinations (see Fig. 1A). This necessitates multiple expert reviews, leading to increasing costs and time.

Related Work: Current FISH image analysis approaches, including machine and deep learning[11,9,29,5], aim at direct spot identification or appearance standardization for easier application of traditional methods like thresholding or gradient techniques [4,15,26]. While certain methods excel with images having clear spot-like signals, they falter with FISH image variability, especially in case of gene amplifications lacking defined spot appearances (see Fig. 1B n=20). In [31] a *RetinaNet*-based dual-network approach has been devised for *HER2* status identification in breast cancer. It classifies nuclei and signals independently, but inaccurately estimates gene copy counts in clusters, defaulting to a fixed number - not necessarily mirroring reality - and inadequately addresses prediction uncertainty. In [12] a two-stream model is introduced by combining whole image and spot information, for FISH image classification, requiring minimal manual annotations, but raising questions about detection reliability and lack of uncertainty quantification. In [14] point clouds were applied to represent localization patterns in synthetic single molecule FISH images effectively, yet this method demands pre-extraction of point positions, limiting its direct applicability to imaging data. To date, no studies have successfully integrated uncertainty modeling into FISH classification.

Contribution: Our work provides four major contributions: (i) An innovative methodology for the precise classification of single-cell FISH images, leveraging synthetic data to bypass the need for costly manual annotations. (ii) A technique for generating the aforementioned synthetic FISH images. (iii) An innovative strategy for embedding inter-class variation into classification through a contrastive learning-based joint objective. (iv) Our approach enables the accurate quantification of classification uncertainty, demonstrating its concordance with human expert judgment and its applicability in the diagnostic processes.

2 Material & Methods

The proposed work consists of two modules. First, we introduce "FISHPainter" a method offering users the flexibility to create FISH images with desired signal characteristics and second, a novel contrastive learning (CL) approach for genetic aberration classification incorporating the uncertainty of classification without manual annotations.

FISHPainter: To address the data sparsity in the field of FISH imaging, especially given its extensive applications across various gene targets and diagnostic use cases, we propose FISHPainter. This tool generates synthetic images, giving users complete control over the modeled FISH modality and data distribution, thereby creating a diverse and well-populated image space, which real-world images, with their prevalence biases, are not sure to provide. Importantly, FISHPainter enables the population of image spaces for cases that lie on the boundaries of different diagnostic categories, which in turn is expected to improve uncertainty estimation in these critical areas.
FISHPainter involves the following steps: (i) Selecting a cell nucleus background from a predefined library to place at the center of an RGB image patch in the blue channel. (ii) Defining positions for signals in the red and green channel, distinguishing between individual and clustered signals — the former randomly placed within the nucleus, and the latter formed by adding signals near a central point, all represented as 2D-Gaussian distributions with user-defined sizes. 3) Applying non-affine transformations to signals to vary appearances. Fig. 1B illustrates generated patches, with n indicating the number of green signals. Implementation available at: https://github.com/SimonBon/FISHPainter

Synthetic Contrastive Learning: Studies show that classification methods based solely on Cross Entropy (CE) inaccurately gauge uncertainty, leading to overconfidence [27,1]. Alternative techniques, including Ensembles [17], Monte-Carlo Dropout [10], and Multi-Head models [19], aim to enhance uncertainty estimation, though they face scalability issues due to their high computational requirements. CL has been recognized for its capacity to incorporate uncertainty into data representations [30,3,16]. Our method, inspired by [28,21], infers prediction uncertainty from softmax outputs of high-quality latent representations of FISH images, jointly training the network's backbone and classification head using CL. Acknowledging that class label-based representations often miss class

variability and therefore inherent uncertainty, our model integrates both class labels and visual similarity into its latent representation, using CE and CL loss. This approach accounts for the levels of intra-class variation by combining NT-Xent [24] and CE in the loss function:

$$\mathcal{L} = -\left(\underbrace{\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)}}_{\text{Contrastive}} + \lambda \underbrace{\left(\sum_{c=1}^{C} y_c \log(\hat{y}_{i,c}) + \sum_{c=1}^{C} y_c \log(\hat{y}_{j,c}) \right)}_{\text{CE for Positive Pair } (i,j)} \right) \tag{1}$$

In Eq. 1, N is the batch size, z_i and z_j are the projected embeddings of two augmented views i and j of a patch, τ is the temperature parameter, and λ is the classification loss weight. C is the class count, y_c and $\hat{y}_{\cdot,c}$ are the target and predicted probabilities for class c, respectively and sim the similarity metric .

In Fig. 2 the model training using the loss function introduced in Eq. 1 is shown: First, for generating input images X and target labels Y with FISHPainter, a set of user-defined signal configurations Q are utilized. The resulting images are augmented with transformations t sampled from a set of transformations T, producing views X_i and X_j. These views are embedded by the network's backbone (blue) into the representation space R. This representation is fed into the projector (orange) to obtain an embedding Z, on which the contrastive loss is calculated, and into the classification head (green), producing a class prediction $\hat{Y}$ for CE loss computation using the target label Y.

Uncertainty: Uncertainty estimation is crucial in medical artificial intelligence applications [23]. Epistemic uncertainty helps to identify out-of-distribution (OOD) samples [28], while aleatoric uncertainty addresses inherent data uncertainties, such as ambiguous signal quality. In this work our main focus lies on aleatoric uncertainty, which we will refer to in the following as uncertainty. In-

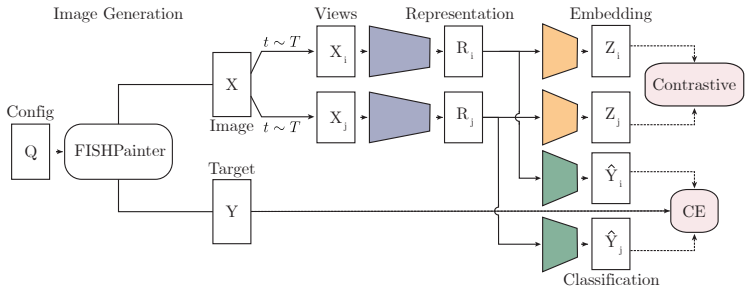

Fig. 2: Configuration of signals Q is input to FISHPainter, producing image X and label Y. Image X is augmented with transformations t sampled from set T to create views X_i and X_j. These views are embedded into representation space R and projected to Z. Contrastive loss is calculated on Z. R is also used by the classification layer to predict $\hat{Y}$, on which CE loss is computed.

spired by [21], we model aleatoric uncertainty $\mathbb{H}_{\text{norm}}(.)$ using normed entropy using Eq. 2:

$$\mathbb{H}_{\text{norm}}(\hat{Y}) = \frac{\mathbb{H}(\hat{Y}) - \mathbb{H}_{\min}(\alpha, C)}{\log(C)} \tag{2}$$

Where $\mathbb{H}(\cdot)$ denotes the entropy function, $\hat{Y}$ represents the softmax output of the model, C is the number of classes, and $\mathbb{H}_{\min}(\alpha, C)$ indicates the minimum entropy for label smoothing value α, which is calculated as follows:

$$\mathbb{H}_{\min}(\alpha, C) = \mathbb{H}\left(\left[1-\alpha, \left(\frac{\alpha}{C-1}\right)_0, \left(\frac{\alpha}{C-1}\right)_1, \ldots, \left(\frac{\alpha}{C-1}\right)_{C-1}\right]\right) \tag{3}$$

3 Experimental Setup

The performance of the proposed method is evaluated on the task of classifying *MYCN* gene aberrations in neuroblastoma [2] with the following three classes: *MYCN Normal* (c_N) - showing 2 green and 2 red signals (2:2 - *MYCN*:*NMI* ratio), *MYCN Gain* (c_G) - showing 3-7 (green) and 2 (red) and *MYCN Amplification* (c_A) with ≥ 8 (green) and 2 (red), as shown in Fig. 1B. We therefore set $C = 3$ in Eq. 2 and Eq. 3.

Data Setup 1: In experiment 1 the generalisation from synthetic FISH data training to real world FISH image classification is analysed using a training dataset of 30,000 synthetic images (10,000 each of c_N, c_G and c_A) utilizing "FISHPainter" (see configurations Q in Supplementary Fig. **??**A). We used a 60% training, 20% validation, and 20% testing split. Additionally, we employ an expert annotated real-world single-cell *MYCN* FISH dataset containing 1814 single-nuclei image patches, obtained from 2 c_N, 2 c_G and 2 c_A cell lines, and use it to assess model generalization in experiment 1, and uncertainty calibration in experiment 3. The following set of augmentations T were chosen to model variations in microscopic images: rotations, flipping, scaling, blurring, channel-specific intensity adjustments, noise, and gradient additions (see parameters and examples in Supplementary Fig. **??**B).

Data Setup 2: In experiment 2 the aim is to assess the method's ability to detect in distribution (ID) and OOD samples. The synthetic training dataset serves as the reference, while the 1814 expert annotated real-world FISH images are considered in-distribution (*Real World ID*). For OOD analysis we use an in-house immunofluorescence dataset (*OOD*), similar yet distinct from FISH (see Fig. **??**C) in targeting cell surface proteins instead of genetic sequences.

Data Setup 3: In experiment 3, we focused on assessing how well the model matches human uncertainty by utilizing annotations of ten experts on a set of 1210 synthetic single-cell FISH images with predefined number of green signals. Images were annotated either c_N, c_G or c_A. Accuracy was measured across all images, while human uncertainty was calculated per image using normalized entropy on the expert annotations.

Method Implementation: We implemented our approach utilizing two sets of augmentations T having different intensity levels: *Ours Heavy*, which applies stronger augmentations, and *Ours Light*, with lighter augmentations (details see Supplementary Fig. **??**B). Both utilize the following parameters: A ResNet-18 [13] backbone modified to reduce the representation space dimension to 128 with a two-layer 64-dimension projection head for CL training [6]. Classification relied on 2 fully connected layers with 128 features, 0.25 dropout, and ReLU activation. The loss function included cosine similarity, 0.05 temperature, 0.5 as λ, and 0.01 as label smoothing (α). We used a Cosine Annealing learning rate scheduler with 5 warmup steps and a 25-epoch cycle, setting learning rate minimum and maximum at 0.00001 and 0.001 respectively, and a batch size (N) of 128.

Baselines: We compared our method against **bigFISH** [15] and a pretrained spot detection **RetinaNet** [18] model from [12], with a classification head fine-tuned on our synthetic training datasets. Additionally, we used three baselines, all employing a **ResNet-18** backbone: (i) **ResNet+CE**: trained using only CE loss, and two baselines using CL with NT-Xent loss and cosine similarity [24], (ii) **CL+Detached**: pretrained with contrastive loss, then froze the backbone and fine-tuned only the classification head with CE, (iii) **CL+Attached**: pretrained with contrastive loss, then fine-tuned both the backbone and the classification head with CE. The set of augmentations T for all baselines can be seen in Fig. **??**B. Fig. **??** shows a schematic illustration of each baseline's training schemes.

4 Results

Experiment 1: The ablation study assesses augmentation effectiveness using classification accuracy across the synthetic test split and the real-world data (Data Setup 1). We compared the use of all augmentations to no augmentations and employed a leave-one-out approach, omitting one augmentation at a time (see Table 1). Augmentations have a minor effect on test split accuracy—94.3% without augmentations versus 94.6% with them. Yet, for real-world datasets, augmentations significantly boost generalization, increasing accuracy from 40.6% to 87.8% without affine transformations, underscoring their critical role in bridging the synthetic and real-world accuracy gap. Gradient and noise augmentations show the greatest impact when ommited (see Table 1 column Grad&Noise).

Experiment 2: The detection of OOD and ID samples is visually evaluated using Data Setup 2. Fig. 3A shows a UMAP [20] of the final latent representation of *Ours Heavy* (see Fig. **??** for other baselines). The real-world dataset significantly overlaps with the synthetic training dataset, showing strong generalization, while the majority of OOD samples occupy a distinct area of the feature space, making them easily separable from ID data. Fig. **??**C shows OOD samples embedded in an ID latent region, illustrating signals that could be mistaken for clustered *MYCN* signals.

	Augmentations		Augmentations Omitted						
	None	All	Affine	Blur	Flip	Grad.	Noise	Int.	Grad&Noise
Synth. Test Set	94.3	94.6	94.7	95.1	95.2	94.7	95.2	**96.0**	94.9
Real World	40.6	86.1	**87.8**	85.4	87.2	80.5	80.8	86.7	45.0

Table 1: Classification accuracies (in %) for the synthetic test set and the manually annotated real-world dataset. Int.: intensity scaling, Grad.: image gradient.

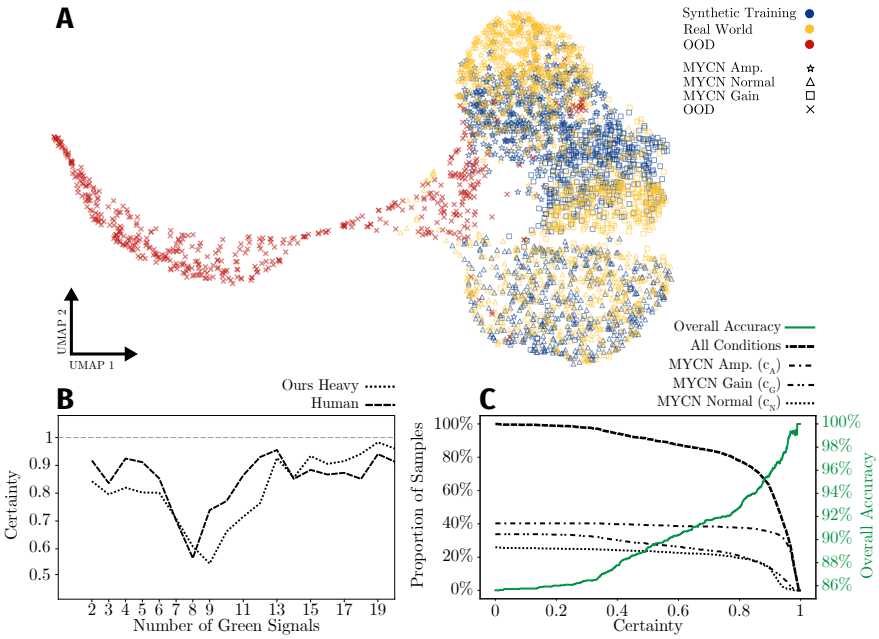

Fig. 3: **A**: 2D latent space visualization of *Ours Heavy* on the synthetic training dataset (blue), the real world test dataset (yellow) and the OOD dataset (red). **B**: Comparison of model to human certainty when classifying synthetic FISH images with set number of green signals. **C**: Accuracy results and the distribution of remaining classes, conditioned on certainty (example: at 0.8 certainty $\sim 78\%$ data remaining with $\sim 93\%$ overall accuracy).

Experiment 3: Accuracy of ten human annotators and our model on Data Setup 1 is presented in Table 2. Our method achieves 90.5% accuracy, close to the top annotator's 90.6% and exceeds the expert average by 2%. Fig. 3B shows the alignment between model-generated aleatoric uncertainty, calculated following Eq. 2 and inter-expert agreement, highlighting similar uncertainty patterns between our model and the annotators, especially around 8 green signals, which is the classification border between c_A and c_G.

Annotator	1	2	3	4	5	6	7	8	9	10	Combined mean ± std	Ours Heavy
Accuracy [%]	89.3	90.3	88.0	84.7	88.3	89.0	89.0	88.9	86.9	**90.6**	88.5 ± 1.6	90.5

Table 2: Comparison of *Ours Heavy* and human annotator accuracy

We assess the Expected Calibration Error (ECE) [8] including positive ECE (overconfidence) and negative ECE (underconfidence). Fig. 4A shows our approach's ECE results using annotated real-world data from Data Setup 1 compared to the baselines. Fig. 4B presents the calibration diagram, featuring a curve indicating data distribution across certainty levels for *Ours Heavy*. The calibration diagram for baseline methods is visualized in Fig. **??**. Our method is well calibrated with the highest overall and negative ECE using heavy augmentations and the lowest positive ECE using light augmentations. In Fig. **??** instances of high model uncertainty are presented highlighting ambiguity, particularly when classifying c_N and c_G (A), where distinguishing between 2 or 3 *MYCN* signals is subjective, and between c_G and c_A (B), where determining the exact *MYCN* copy number is challenging. This affirms that model-generated- and human- uncertainty align, given that these instances are inherently ambiguous and their classification ultimately relies on subjective judgment.
Fig. 3C shows the accuracy and class distribution for *Ours Heavy* conditioned on certainty thresholds, excluding samples below the threshold. Initially, uncertain samples from c_G are removed, followed by uncertain c_N samples, and finally c_A cases. This indicates that c_A cases are easier to classify due to their distinct visual differences from c_G or c_N (compare n=20 with n=2 and n=5 in Fig. 1B). Conditioning plots for all baselines are shown in Fig. **??**.
Table 3 shows accuracy scores for each baseline at different remaining sample percentages, conditioned on the yielded aleatoric certainty, calculated following Eq. 2. Since *bigFISH* and *RetinaNet* do not provide uncertainties, their accuracy is measured only at 100%. Accuracy values indicate they cannot compete with models that do not rely on spot signals. The *CL+Attached* model performs well with 88.3% accuracy at 100% data but becomes unstable for higher certainty values (see 20% - 5%). Notably, *Ours Light* shows 90.3% accuracy at 95% data retention, and *Ours Heavy* achieves 98.9% accuracy at 30% of the most certain samples, further increasing to 100% at 5%. *ResNet+CE* exhibits overconfidence, reaching maximum certainty for the top 40% of samples but only achieving 95.6% accuracy, highlighting its overconfidence, which is also observable in ECE plot of Fig. **??**A. In contrast, *Ours Heavy* improves accuracy as the conditioning becomes more stringent. We hypothesize that the performance difference between *Ours Heavy* and *Ours Light* is due to a more restricted feature space under the contrastive loss in *Ours Heavy*. This leads to poorer cluster separation in the embedding space, but better calibration, particularly for positive ECE and overconfidence. Additionally, we observe that lighter augmentation intensity im-

A

Model	ECE ↓	Pos. ECE ↓	Neg. ECE ↓
ResNet + CE	7.05	7.85	7.00
CL + Detached	9.59	9.78	6.61.
CL + Attached	2.71	5.18	1.89
Ours Light	3.51	**2.49**	4.46
Ours Heavy	**2.69**	4.05	**1.68**

all values $\times 10^{-2}$

B

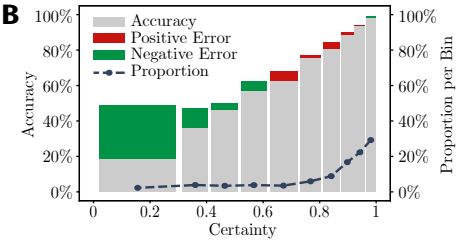

Fig. 4: **A**: ECE, positive ECE, and negative ECE comparison between our method and all baseline approaches on real world data. **B**: Calibration chart for *Ours Heavy*, showing the sample distribution across certainty bins.

proves class cluster separation in the embedding space, but this results in poorer calibration by not adequately representing the transition between classes.

		Percentage of remaining data										
Model Type		100%	95%	90%	75%	50%	40%	30%	20%	15%	10%	5%
bigFISH [15]	Accuracies [%]	59,4	not applicable									
RetinaNet [12]		71,9	not applicable									
ResNet+CE		86,4	88,5	90,8	94,7	95,6	95,6	-	-	-	-	-
CL+Detached		73	73,8	75,3	80,4	89	93,3	96,2	98,1	97,8	97,2	98,9
CL+Attached		**88,3**	90,1	91,4	94,1	96,3	96,7	96,7	95,6	94,7	92,7	93,1
Ours Light		**88,3**	**90,3**	**91,7**	**95,2**	96,0	96,4	96,2	96,9	98,2	97,9	100
Ours Heavy		85,6	87,4	89,6	93,8	**96,7**	**97,5**	**98,9**	**99,2**	**99,3**	**99,4**	**100**

Table 3: Accuracy comparisons between our method and baseline models after excluding the most uncertain samples, based on conditioning on certainty. "-" indicates that all data has been removed, so accuracy can not be calculated.

5 Conclusion

We introduced an adaptable, end-to-end methodology for classifying FISH single-cell images without manual annotations using "FISHPainter", which only requires a clear class definition. A key aspect of our approach is the use of synthetic data for training, enabling us to densely model the image space of FISH images. Our results demonstrate that FISH-specific augmentations are essential for transitioning from synthetic training to real-world test settings. By using synthetic FISH images with appropriate augmentations and incorporating cross-entropy into the contrastive loss, our method produces a well-calibrated model. This integration of uncertainty is crucial for advancing digital pathology diagnostics, providing more reliable and interpretable results.

Acknowledgments. This research was supported by Vienna Science and Technology Fund (WWTF) PREDICTOME [10.47379/LS20065], EU EUCAIM (No.101100633-EUCAIM) and the Austrian Science Fund (FWF) MAPMET [10.55776/P35841].

References

1. Algan, G., Ulusoy, I.: Image classification with deep learning in the presence of noisy labels: A survey. Knowledge-Based Systems **215** (3 2021)
2. Ambros, P.F., Ambros, I.M., Brodeur, G.M., Haber, M., Khan, J., Nakagawara, A., Schleiermacher, G., Speleman, F., Spitz, R., London, W.B., Cohn, S.L., Pearson, A.D., Maris, J.M.: International consensus for neuroblastoma molecular diagnostics: Report from the international neuroblastoma risk group (inrg) biology committee (5 2009)
3. Ardeshir, S., Azizan, N.: Uncertainty in contrastive learning: On the predictability of downstream performance (7 2022)
4. Bahry, E., Breimann, L., Zouinkhi, M., Epstein, L., Kolyvanov, K., Long, X., Harrington, K.I.S., Lionnet, T., Preibisch, S.: Rs-fish: Precise, interactive, fast, and scalable fish spot detection. bioRxiv p. 2021.03.09.434205 (10 2021)
5. Bouilhol, E., Lefevre, E., Dartigues, B., Brackin, R., Savulescu, A.F., Nikolski, M.: Deepspot: a deep neural network for rna spot enhancement in smfish microscopy images (2021)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020)
7. Cohn, S.L., Pearson, A.D., London, W.B., Monclair, T., Ambros, P.F., Brodeur, G.M., Faldum, A., Hero, B., Iehara, T., Machin, D., Mosseri, V., Simon, T., Garaventa, A., Castel, V., Matthay, K.K.: The international neuroblastoma risk group (inrg) classification system: An inrg task force report. Journal of Clinical Oncology **27**, 289–297 (1 2009)
8. Ding, Y., Liu, J., Xiong, J., Shi, Y.: Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops **2020-June**, 22–31 (3 2019)
9. Eichenberger, B.T., Zhan, Y., Rempfler, M., Giorgetti, L., Chao, J.A.: Deepblink: Threshold-independent detection and localization of diffraction-limited spots. Nucleic Acids Research **49**, 7292–7297 (7 2021)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning (6 2015)
11. Gudla, P.R., Nakayama, K., Pegoraro, G., Misteli, T.: Spotlearn: Convolutional neural network for detection of fluorescence in situ hybridization (fish) signals in high-throughput imaging approaches. Cold Spring Harbor symposia on quantitative biology **82**, 57–70 (2017)
12. Gutwein, S., Kampel, M., Sabine, T.M., Licandro, R.: Genuine: Genomic and nucleus information embedding for single cell genetic alteration classification in microscopic images. Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (2024)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2016-December**, 770–778 (12 2015)

14. Imbert, A., Mueller, F., Walter, T.: Pointfish – learning point cloud representations for rna localization patterns (2 2023)
15. Imbert, A., Ouyang, W., Safieddine, A., Coleno, E., Zimmer, C., Bertrand, E., Walter, T., Mueller, F.: Fish-quant v2: a scalable and modular tool for smfish image analysis. RNA **28**, 786–795 (6 2022)
16. Kirchhof, M., Kasneci, E., Oh, S.J.: Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs (2 2023)
17. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles (12 2016)
18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 318–327 (8 2017)
19. Linmans, J., van der Laak, J., Litjens, G.: Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks (1 2020)
20. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [cs, stat] (2020), comment: Reference implementation available at http://github.com/lmcinnes/umap
21. Mukhoti, J., Kirsch, A., Amersfoort, J.V., Torr, P.H.S., Gal, Y.: Deep deterministic uncertainty: A new simple baseline (2023)
22. Penault-Llorca, F., Bilous, M., Dowsett, M., Hanna, W., Osamura, R.Y., Rüschoff, J., Vijver, M.V.D.: Emerging technologies for assessing her2 amplification. American Journal of Clinical Pathology **132**, 539–548 (10 2009)
23. Seoni, S., Jahmunah, V., Salvi, M., Barua, P.D., Molinari, F., Acharya, U.R.: Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023) (10 2023)
24. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in Neural Information Processing Systems **29** (2016)
25. Tang, Z., Wang, L., Tang, G., Medeiros, L.J.: Fluorescence in situ hybridization (fish) for detecting anaplastic lymphoma kinase (alk) rearrangement in lung cancer: Clinically relevant technical aspects (8 2019)
26. Tinevez, J.Y., Perry, N., Schindelin, J., Hoopes, G.M., Reynolds, G.D., Laplantine, E., Bednarek, S.Y., Shorte, S.L., Eliceiri, K.W.: Trackmate: An open and extensible platform for single-particle tracking. Methods **115**, 80–90 (2 2017)
27. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization (5 2022)
28. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., Cemgil, T., Eslami, S.M.A., Ronneberger, O.: Contrastive training for improved out-of-distribution detection (7 2020)
29. Wollmann, T., Ritter, C., Dohrke, J.N., Lee, J.Y., Bartenschlager, R., Rohr, K.: Detnet: Deep neural network for particle detection in fluorescence microscopy images. Proceedings - International Symposium on Biomedical Imaging **2019-April**, 517–520 (4 2019)
30. Wu, M., Goodman, N.: A simple framework for uncertainty in contrastive learning (10 2020)
31. Zakrzewski, F., de Back, W., Weigert, M., Wenke, T., Zeugner, S., Mantey, R., Sperling, C., Friedrich, K., Roeder, I., Aust, D., Baretton, G., Hönscheid, P.: Automated detection of the her2 gene amplification status in fluorescence in situ hybridization images for the diagnostics of cancer tissues. Scientific Reports **9** (12 2019)