Investigating Imaging, Annotation and Self-Supervision for the Classification of Continuously Developing Cells in Histological Whole Slide Images

Sebastian Thiele[*,1,2] Jacqueline Kockwelp[*,1,2,3]
Joachim Wistuba[3] Sabine Kliesch[3] Jörg Gromoll[3] Benjamin Risse[†1,2]
[1]Institute for Geoinformatics, University of Münster
[2]Faculty of Mathematics and Computer Science, University of Münster
[3]Centre of Reproductive Medicine and Andrology, University Hospital Münster
firstname.lastname@uni-muenster.de[1,2] firstname.lastname@ukmuenster.de[3]

Abstract

The analysis of individual cells is increasingly automated through deep learning techniques. This is particularly relevant for high-resolution whole slide images (WSIs), which can contain thousands of cells, making manual evaluation impractical. This increase in automation, however, requires higher levels of standardisation (with respect to the scanning hardware, settings and staining) and is further aggravated by the dynamics of the underlying cellular processes, rendering unique cell classifications difficult. To address these difficulties we investigated the entire processing pipeline (from imaging over annotation to model training) and study its underlying trade-offs. In particular, we created a new dataset comprising of more than $6,300$ *labelled and* $500,000$ *unlabelled cells scanned using two different scan settings, resulting in fully registered image pairs with varying level of detail and quality. Using these alternative dataset versions we analysed the impact of inter- and intra-variability between three different annotators and addressed the challenge of limited labelled data by comparing the impact of different self-supervised pretraining strategies. Overall, our analyses provide new insights into the dependencies between imaging, annotation, self-supervision and deep learning-based classification, especially in the context of continuously developing cells and demonstrate the beneficial impact of these considerations on the overall classification accuracy. Code is available at* https://zivgitlab.uni-muenster.de/cvmls/icdc *and the data will be shared upon qualified request due to data privacy laws.*

1. Introduction

In recent years, the application of deep learning-based methods for the analysis of high-resolution histological whole slide images (WSIs) has significantly increased. These analyses extend down to the cellular level, where individual cells are extracted and classified [22, 23, 36, 37]. However, a key challenge in classification is that WSIs provide only a snapshot in time, capturing cells within a highly dynamic developmental system. As a result, cells may be in transitional states between developmental stages, making it challenging to classify them into distinct categories, even for experienced histologists. This has a variety of consequences and implications for the involved imaging, annotation and analysis tasks as shown in Figure 1. For example, ambiguous labels might be the result of the aleatoric uncertainty intrinsically present in continuously developing cells aggravating the performance of supervised machine learning methods due to errors in the ground truth labels. We chose to use histological whole slide scan imagery depicting human testis tissue for this study, because these tissue sections provide a cross-sectional view of the seminiferous tubules, where the spermatogenesis takes place. Spermatogenesis entails a dynamic interplay of mitotic and meiotic processes, intricately intertwined with a complex pattern of germ cell differentiation, hence providing dynamic and often ambiguous cell stages within one tissue and in close spatial proximity [33] and therefore making the task of classifying cells during spermatogenisis a suitable basis for our analysis. Figure 2a visually represents this complexity, illustrating the progression from the periphery to the centre of the tubules. We identify eight distinct cell types involved in spermatogenesis: peritubular myotic cell (**PMC**), Sertoli cell (**SC**), spermatogonium A dark (**Spg Ad**), spermatogonium (**Spg**), premeiotic spermatocyte (**Pl-Z**), pachytene spermatocyte (**P**), round spermatid (**rsptd**), and elongated spermatid (**esptd**) (see Figure 2b). Notably, PMC and SC

[*]These authors contributed equally to this work
[†]Corresponding author

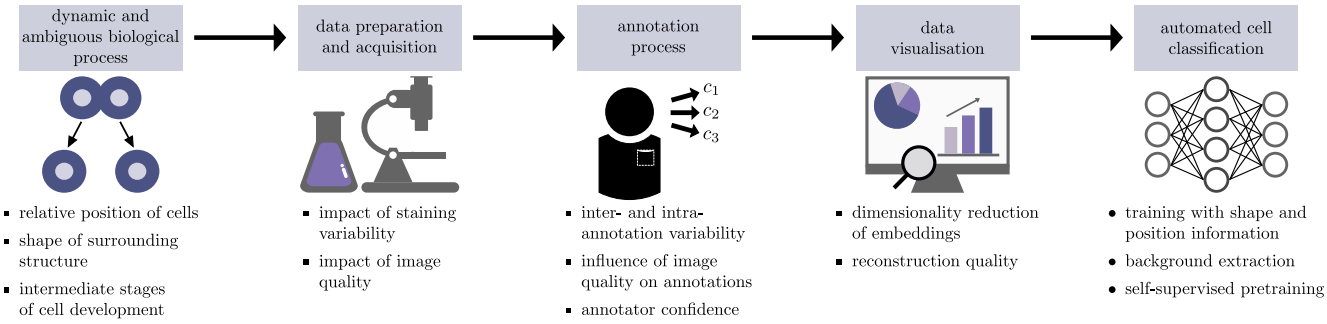


Figure 1. Motivation and Overview. Involved steps (top row), illustration (mid row) and technical challenges (bottom row) are indicated. Our analysis addresses the mentioned challenges and discusses the mutual impact of the steps on the overall reliability for cell classification.

are somatic cells and thus differ from the germ cells that undergo development in stages.

Figure 1 visualises all of our analyses and the insights gained at each step. First, we analysed the continuous and dynamic biological process of cell differentiation and manual cell stage classification by collaborating with experienced biologists. During this analysis, we identified three properties that biologists found beneficial for manual classification, namely the relative position of the cell within the tubule, the tubule's shape, and the background information. These properties were integrated and evaluated as additional information for the neural network. Next, we addressed the challenge of varying data quality in whole slide imaging. Different laboratories use various whole slide scanners and scan settings, which significantly impact the image quality. We investigated the impact of the data quality on the performance of neural networks by scanning biopsy samples with two different scan settings, namely the standard settings (typically used in medical practice) and an alternative setting specifically optimised for higher resolution and better focused images. This approach resulted in two datasets capturing the same areas of interest under different scan conditions. For optimal comparability, both datasets were registered to each other. Additionally, in many laboratories, tissue sections are stained manually, which introduces the risk of colour variability due to manual biases. Figure 2c presents a qualitative comparison of the used scan settings and different staining intensities. In the third analysis step, we examined the annotation process. Three human annotators with greatly varying experience levels labelled the same cells in images of different qualities. We analysed inter- and intra-variability and the influence of scan settings on annotation. In a final step we visualised three self-supervised learning (SSL) techniques and analysed their impact as pretraining methods for our task. In an extensive evaluation, we pretrained six commonly used backbones of varying types, namely ResNet [16] and Vision Transformers (ViT) [10], and sizes using three self-supervised pretraining methods - one of which could only be applied to ViTs - across two scan settings, yielding a total of 30 pretrained models. These models, along with a baseline that was either trained from scratch (ResNets and Vision Transformers) or pretrained on ImageNet (ResNets), were subjected to 5-fold cross-validation across various training parameter configurations, resulting in 660 models trained in 132 different configurations.

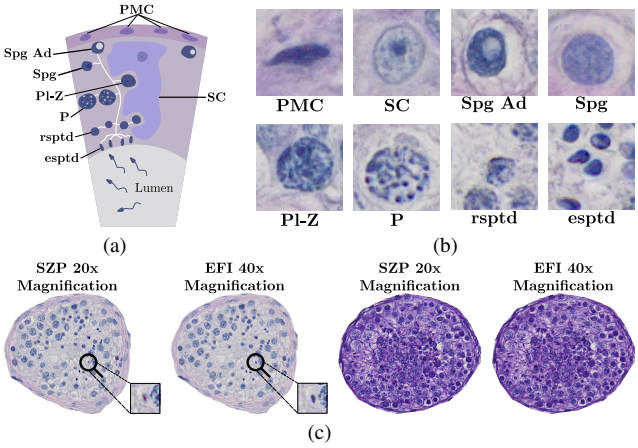


Figure 2. (a) Schematic of spermatogenesis, depicting the stages of male gamete production: differentiation of spermatogonia into spermatocytes, meiotic division yielding four spermatids, and maturation of spermatids into spermatozoa. (b) Examples of the eight nucleus classes utilised in this work. (c) Comparison between the two utilised scan profiles (Single Z-Plane (SZP) and Extended Focus Imaging (EFI)) and demonstration of staining differences caused by manual biases.

2. Related Work

With the advent of deep learning algorithms, numerous studies have integrated these techniques into the automatic classification of cells in histological slides [34, 38]. It is often necessary to first detect and extract cells from the images. One approach is to use end-to-end models for simultaneous cell detection and classification [35]. Alternatively,

deep learning models can be integrated processing pipelines for cell image extraction which often also use classical (intermediate) image processing and analysis routines [22,23]. With Cellpose [29, 36] a foundation model for cell-type agnostic segmentation was published, that enables retraining on specific datasets and therefore serves as a versatile model for both, end-to-end and pipeline-based approaches. However, research on the classification of continuously developing cells in WSIs remains limited.

In parallel to the developments in deep learning, whole slide scan imagery has become increasingly popular for the digitisation in pathology [30]. Being a cost-effective technology for automatised high-throughput data generation these images are frequently used as the input for automatic image analysis algorithms [8, 19]. However, histological slides can be very heterogeneous. For example, different staining protocols or section slicing thicknesses can lead to severe differences in WSIs. Moreover, varying digitisation processes can introduce artifacts such as colour and contrast variations or out-of-focus regions. These artifacts can lead to ambiguous annotations, diagnostic errors and adversely affect the performance of automated analysis algorithms [24,26]. As a consequence, large amounts of consistently annotated data, especially for rare pathologies, are still scarce, which further aggravates the utilisation of deep learning models for WSI analyses.

To overcome this limitation self-supervised learning techniques appear to be a promising candidate. While self-supervised learning had its first breakthrough in the context of language models [3, 9], similar techniques were successfully adapted to images. Contrastive methods like SimCLR [6] or DINO [5] emerged, that rely on comparing heavily augmented image views during training to learn suitable representations. Reconstructive approaches on the other hand are tasked to restore an image from an augmented version. For example, Masked Autoencoders (MAE) [14] aim to reconstruct an image that has been masked with a high masking ratio, while AIM [11] tries to predict the correct order of randomly shuffled patches of an image. The applicability of these techniques has also been studied in the context of medical data and histopathology [1,18,20]. However, literature on self-supervision for cell classification is still underrepresented.

3. Method

3.1. Dataset Acquisition

All testis biopsies were bouin-fixed, paraffin embedded, stained with Periodic Acid-Schiff (PAS) [2] and scanned by using an Olympus VS120 slide scanner in two different scan settings, namely the standard and an optimised setting. In the standard setting, from now on referred to as **20× SZP**, a single z-plane was scanned at $20\times$ magnification. This setting is commonly used in everyday medical practice to save time and storage space. In the optimised setting, multiple z-planes were scanned at an increased $40\times$ magnification while utilising the Extended Focus Imaging (EFI) function. The EFI function captures images of samples that extend beyond the depth of focus of the objective and combines them into a single, fully focused image. This optimised setting is hereafter referred to as **40× EFI**. Note that the $40\times$ EFI setting not only increases resolution but also enhances overall image quality by providing improved focus compared to $20\times$ SZP (see Figure 2c).

We used 68 WSIs from 21 different patients to create our datasets. Ethical approval was obtained (Ethics Committee of the Medical Faculty of Münster and State Medical Board no. 2008-090-f-S), and all participants provided written informed consent. The scans had an average size of $10,421 \times 10,229$ pixels (range, $4,997 - 20,851 \times 4,291 - 18,887$) using $20\times$ SZP setting and $23,126 \times 23,422$ pixels (range, $11,039 - 45,483 \times 8,666 - 48,618$) using $40\times$ EFI setting. Each pixel had a physical size of $0.34 \times 0.34 \mu m^2$ for $20\times$ SZP and $0.17 \times 0.17 \mu m^2$ for $40\times$ EFI. Storage requirements for the scans in PNG format were approximately 9.4 GB ($20\times$ SZP) or 39.6 GB ($40\times$ EFI).

Data Preprocessing for Scan Quality Analysis Initially, we converted the original VSI scan files into PNG files using Bio-Formats [25]. Subsequently, we segmented all nuclei contained within the tubules using the *cyto* model from the Cellpose model zoo [29, 36] that was fine-tuned using $25,935$ segmented nuclei from out specific data domain. The WSIs were initially annotated by an annotator with three years of experience in testicular histology (referred to as Annotator 1) using the $40\times$ EFI scan settings, according to the previously mentioned eight cell classes. To assess the impact of different scan settings and resulting image quality, such as focus, on performance, these annotations were automatically mapped to the $20\times$ SZP WSIs. Due to manual selection of scanned regions, direct transfer of annotations was not feasible. Instead, we calculated homography matrices to facilitate the mapping between corresponding image pairs. To achieve this, we employed ORB features [32] to detect and compute keypoints and descriptors. Matches were calculated using the Hamming distance based on these descriptors. Subsequently, we used the matched keypoints to compute the homography matrix using RANSAC [12] with a threshold of 5.0. Afterwards, we cropped images to sizes of 40×40 pixels or 80×80 pixels, corresponding to the scan settings for each annotated nucleus. Additionally, for each cell, we extracted a segmented version, in which only the nucleus of interest is visible and everything else, including other cells, is masked. In our experimental Section 4, we refer to this difference as trained with or without background.

	PMC	Spg Ad	Spg	Pl-Z	P	rsptd	esptd	SC	all
train/val	355	231	459	826	1,097	665	951	616	5,200
test	103	48	121	132	224	184	197	142	1,151
unlabeled	-	-	-	-	-	-	-	-	503,189

Table 1. Overview of our datasets.

Resulting Datasets In total, we created three distinct datasets for this study. The first train/validation dataset comprises $5,200$ single cell images and was used for all cross-validation experiments. The second dataset was used for testing and contains $1,151$ cells from a separate WSI. The third dataset consists of $503,189$ cells extracted using the fine-tuned Cellpose model. This third dataset is separate from the train/validation and test datasets and was used to study the impact of self-supervised pretraining. A detailed overview of our datasets is given in Table 1.

3.2. Integrating Additional Bio-Inspired Variables

In an initial qualitative assessment of the underlying cell differentiation process we collaborated with domain experts and investigated their strategies when analysing cells from testicular sections. These analyses revealed three important features used for manual cell stage classification, namely the relative position of the cell within the tubule, the tubule's overall shape, and the cell background information. To incorporate the cell's localisation, we calculated the minimum distance from a nucleus to the corresponding tubule outer wall as an additional input feature for the neural network (see Figure 3). Note that this distance has a direct physiological interpretation since cell differentiation of spermatogenesis is organised from the tubule's outer wall to the centre (lumen; see Figure 2). Given that annotated tubules in our dataset vary in shape and size, we normalised the distance value using $\frac{\text{minimal bounding box width}}{2*\text{compactness}}$ as a factor. The *minimal bounding box width* refers to the width of the smallest bounding box that can enclose the tubule. The term $compactness = 4\pi\frac{Area}{Perimeter^2}$ [31], which serves as a measure of the roundness of the tubule, provides insights into the incision type. The impact of background information is evaluated by training networks with and without masking the cell neighbourhood using cell segmentations.

Metrics Reflecting Intermediate Cell Stages In consideration of the dynamic nature of cell development, where cells may exist in intermediate stages complicating precise classification, we used additional metrics to assess classification accuracy beyond standard Top-1 accuracy (**Top-1 Acc**) to provide a more nuanced evaluation. Specifically, we used the following metrics:

Top-2 Acc: considers a prediction as correct if the true class appears within the top two predicted probabilities.

Adjacent-1 Acc: correct classification includes not only exact matches but also predictions of predecessor or successor classes in the developmental sequence, if applicable. Cells with no intermediate developmental stages (PMC and SC) must be predicted exactly to be considered correct.

Adjacent-1 Dynamic Acc: Similar to Adjacent-1 Acc, but restricted to evaluating accuracy only among cell labels that undergo intermediate developmental stages.

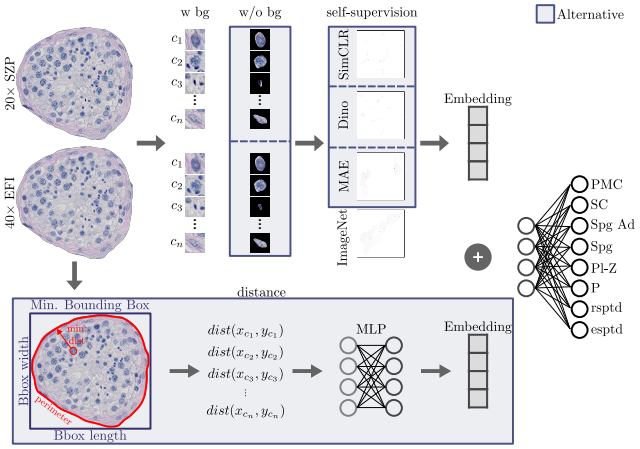


Figure 3. From the $20\times$ SZP setting and $40\times$ EFI WSIs, individual cell crops are extracted, either with (w bg) or without (w/o bg) background. A classification network is trained on the resulting single-cell dataset. Three self-supervised approaches are evaluated. The resulting embedding is optionally combined with the extracted and compactness-normalised minimal distance information of the cell, then classified according to eight class labels.

3.3. Self-Supervised Learning

In recent years it has been shown that self-supervised learning methods can increase a model's performance in several medical tasks, especially in scenarios such as ours where there is an abundance of raw image data, but only a small amount of labels [1, 18, 20]. We analysed the impact of three frequently used yet technically different self-supervision approaches, namely SimCLR [6], DINO [5] and MAE [14], on our dynamic cell classification task.

SimCLR [6] is a contrastive learning technique in which pairs of augmentations of an image are compared to themselves and all other samples in the current batch. The representation similarity between the augmentations of the same image is maximised and the similarity to any other sample is minimised to prevent representation collapse.

DINO [5] is often considered to be a contrastive self-supervised learning technique, although it does not compare samples with counterexamples from the current training batch. Instead it combines findings from other previous works such as SwAV [4], BYOL [13] and MoCo [15] into a knowledge distillation [17] framework with no labels. DINO generates several global and local augmented views

of the same image. It feeds the global views to a frozen teacher network, whose parameters are set to the exponential moving average of the student network, which is fed with the local views. The student is then trained by aligning its output with the centred and sharpened output of the teacher network, thereby avoiding representation collapse. It is therefore not necessary to compare to negatives sampled from the same training batch.

MAE [14] was introduced to exploit properties of Vision Transformers during a reconstruction-based self-supervision approach and is therefore not considered a contrastive learning technique. During training, random patches of the images are masked and the remaining patches are fed into a transformer encoder. Afterwards a smaller transformer decoder aims to reconstruct the original image by filling in token representations of missing image patches. Since the encoder does not consider tokens of missing patches and the decoder can be small this framework allows for very efficient pretraining.

4. Experiments

4.1. Analysis of Annotation Process

To analyse the reliability of our ground truth data regarding the six dynamic cell classes and the annotator confidence, we selected eight cells from each class of the test dataset and had them labelled by three different annotators in their respective $20\times$ SZP and $40\times$ EFI versions. We ensured that exactly the same cell images were labelled separately by each annotator. To account for different levels of experience we selected the following annotators: Annotator 1 (previously mentioned with three years of experience in working with testicular histology WSIs); Annotator 2 (more than 30 years of experience as a histologist routinely working with scans with standard scan settings); Annotator 3 (no prior experience, using reference images). For evaluating annotator confidence, we introduced intermediate labels in addition to the six dynamic class labels. Annotators could use these intermediate labels when they believed a cell was between two stages, resulting in a total of 11 label options for this experiment. During the labelling process, we used 40×40 px ($20\times$ SZP setting) or 80×80 px ($40\times$ EFI setting) image crops with background. No additional localisation information was given to ensure annotations based on the cell appearance and close neighbourhood only.

		20x SZP samples			40x EFI samples		
		A1	A2	A3	A1	A2	A3
A1	Top-1	45.8	56.2	64.5	66.6	64.5	52.0
	Adj.-1 Dyn.	93.7	83.3	81.2	95.8	87.5	89.5
	Adj.-2 Dyn.	97.9	93.7	87.5	1.0	93.7	95.8
A2	Top-1	-	-	50.0	-	-	50.0
	Adj.-1 Dyn.			70.8			81.2
	Adj.-2 Dyn.			83.3			93.7

Table 2. Inter- and intra-variability for samples in different scan qualities.

	Top-1	Adj.-1 Dyn.	Adj.-2 Dyn.
A1	60.4	93.7	1.0
A2	58.3	79.1	89.5
A3	50.0	89.5	93.7

Table 3. Scan quality Intra-variability.

Table 2 presents the results for intra-variability, measured for Annotator 1, who annotated the data twice, and inter-variability among all three annotators. We used the following metrics: Top-1 Acc, the previously described Adjacent-1 Dynamic (**Adj.-1 Dyn.**) Acc (which includes the classes immediately before and after the correct class), and an extended Adjacent-2 Dynamic (**Adj.-2 Dyn.**) Acc (which includes two classes before and after), similar to the previous Adjacent-1 Dynamic Acc on six classes due to the introduction of intermediate labels. The investigation of intra-variability reveals significant differences within the labelling process, with a Top-1 Acc of 45.8% for the $20\times$ SZP data and 66.6% for the $40\times$ EFI data. Inter-variability, measured across all annotators, also indicates noticeable differences and uncertainties, though these are less pronounced for the $40\times$ EFI dataset. The average Adjacent-2 Acc for the $20\times$ SZP data is 88.1%, while for the $40\times$ EFI dataset, it is 94.4%. Overall the results indicate that despite the experience of the annotator, it is impossible to provide unambiguous labels due to the complexity of the task.

Additionally, we examined the intra-variability concerning the different image qualities for each annotator (see Table 3 and Figure A.1). This analysis shows significant differences in cell evaluations depending on the scan settings. The Top-1 Acc ranges only from 50% to a maximum of 60.4% and even for a highly experienced testis histologist, the difference in Adjacent-2 Acc exceeds 10%. This analysis further demonstrates that the quality of the data has a substantial impact.

4.2. Influence of Additional Bio-Inspired Variables and Scan Settings

In our initial baseline experiments, we trained six different network architectures and analysed the impact of incorporating distance information as an additional input, varying levels of scan quality, the effect of masking background information in cell images, and the influence of ImageNet pretraining [7]. For all experiments, we employed a 5-fold cross-validation (see Table A.1).

The ResNet $(18, 50, 101)$ and Vision Transformer (T,S,B) architectures were used for all experiments. In our experiments we chose a Vision Transformer patch size of 8 for the $40\times$ EFI and a patch size 4 for the $20\times$ SZP WSIs. Each token therefore covers the same real world area. We analysed the impact of out-of-domain pretraining by comparing ImageNet pretrained versions of the ResNet variants with the models trained from scratch. While pretrained ViT models exist, the specific image and patch size combinations we utilised are not available. Consequently, we trained the ViTs from scratch for this experiment. Input images were normalised with channel mean and standard deviation of the respective fold's training data to maintain consistency across the dataset. To enhance the model's ro-

bustness and generalisation, we applied random horizontal and vertical flip, random rotation with a maximum angle of 180 degrees and random translation up to 1% along both axes. If applicable, we employed distance information augmentation by setting the distance values randomly to -0.5 with a probability of 10%. Training was carried out with the AdamW [27] optimiser, a learning rate scheduler starting at 10^{-3} with a reducing factor of 0.5 and a patience of 10 and a batch size of 128, over a total of 100 epochs. The ViT-S and B models started out with a lower learning rate of 10^{-4}.

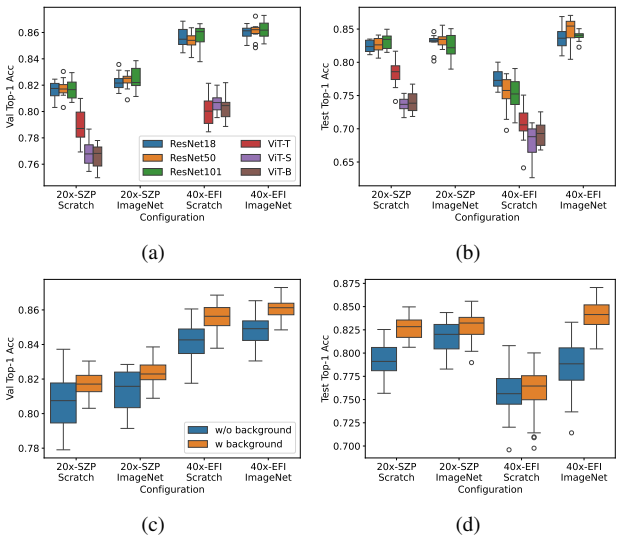


Figure 4. Top-1 Acc of different backbones and scan settings: val (a) , test (b); only with background; with and without distance information (180 models in 36 distinct configurations). The legend of (a) also applies to (b). Comparison of Top-1 Acc training with and without background in both scan settings: val (c), test (d); only ResNets; with and without distance information (240 models in 48 distinct configurations). The legend of (c) also applies to (d).

All aggregated Top-1 Acc, Top-2 Acc, Adjacent-1 Acc and Adjacent-1 Dynamic Acc values for validation and test set are presented in detail in Tables A.5 (from scratch) and A.6 (ImageNet). Figures 4a and 4b demonstrate that Vision Transformers generally perform worse when trained from scratch compared to ResNets, which is consistent with current research given the relatively small size of the labelled portion of our dataset [21]. ResNets pretrained on ImageNet outperform those trained from scratch, albeit by a smaller margin. Utilising the enhanced scan settings of $40\times$ magnification and EFI resulted in a performance increase on validation and test, except for the networks trained from scratch, whose test performance declined (see Figure 4b).

Background information appears to be generally beneficial for the classification task (see Figures 4c and 4d). This is likely because other cells visible in the background provide additional contextual information in uncertain cases. Incorporating distance information generally improves performance, though some of the best performing configurations on test are achieved without using distance information (see Figures 6a and 6b). We attribute this to the relatively small size of the labelled dataset, which can introduce slight biases that are not always advantageous in the test set.

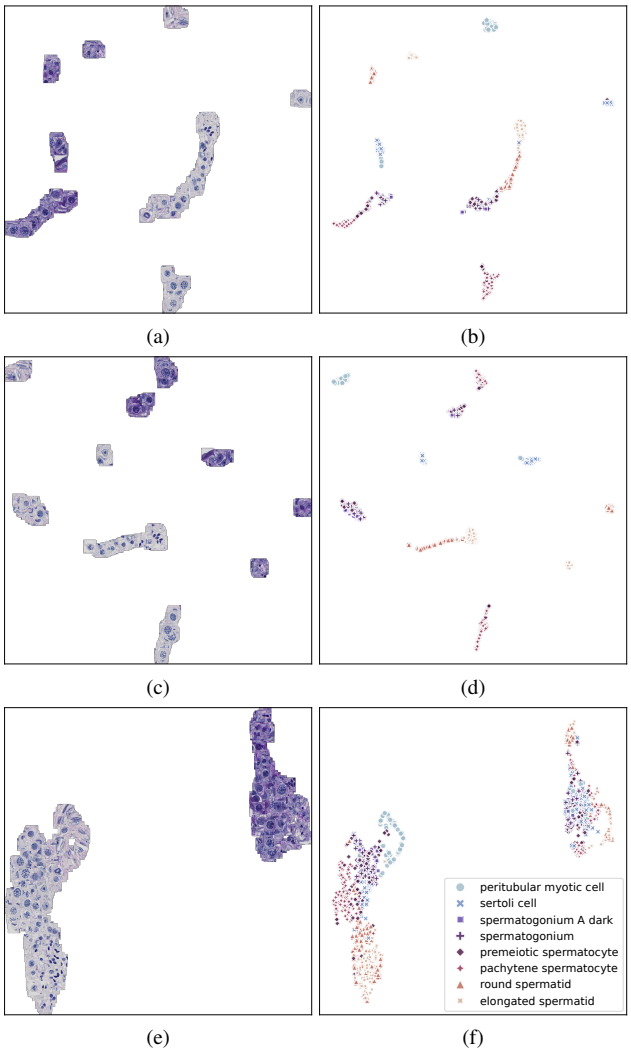
4.3. Self-Supervised Pretraining

We chose the same backbones described in 4.2 for SimCLR and DINO. MAE is a technique entirely based on the transformer architecture. We therefore only chose the ViT variants (T,S,B) as the backbones for this approach. All self-supervised algorithms were trained with similar memory requirements on the nucleus dataset without labels using the AdamW [27] optimiser and validated on the training/validation dataset (see Table 1). The learning rate was reduced by a factor of 0.25 when reaching a plateau in the validation loss (patience 10 epochs). The model with the best validation loss after 200 epochs was chosen as the final pretrained model for any given configuration.

The augmentation settings for the different SSL approaches can be found in A.2.1 (SimCLR), A.2.2 (DINO) and A.2.3 (MAE), while batch sizes are displayed in Tables A.2 (SimCLR), A.3 (DINO) and A.4 (MAE). The initial learning rates for training were set to either 10^{-3} (SimCLR, DINO) or $1.5 * 10^{-4}$ (MAE).

Visualisation We visualised the ViT-B embedding of the self-supervised pretraining on the $40\times$ EFI images (with background) in Figure 5 using UMAP [28]. As previous work suggests [14], the contrastive approaches DINO and SimCLR lead to a clearer separation in the embedding space compared to MAE. However, all approaches have difficulties to cluster the same cell type with vastly different staining intensities, indicating an impact on the classification task. Figure A.2 shows reconstructions of ViT-B MAE on $20\times$ SZP and $40\times$ EFI images. The reconstructions closely resemble the original despite the high masking ratio, with no artificial artefacts being visible in our evaluations.

Fine-Tuning Building on the baseline results shown in Tables A.5 and A.6 our fine-tuning analysis focuses solely on images with background information. All pretrained models were fine-tuned following the training/validation/test split and augmentations described in Sections 4.2. Each model was trained for 30 epochs with AdamW optimiser, a batch size of 128 and a learning rate of 10^{-4}. The pretrained backbones were frozen for the first epoch and afterwards unfrozen with an initial learning rate of 10^{-6}. The learning rate for the backbone was exponentially increased by a factor of 2 until it reached the learning rate of the classifier and kept at 10^{-4} until epoch 22, after which it was exponentially reduced again by a factor of 0.5 each epoch until the end of the training.



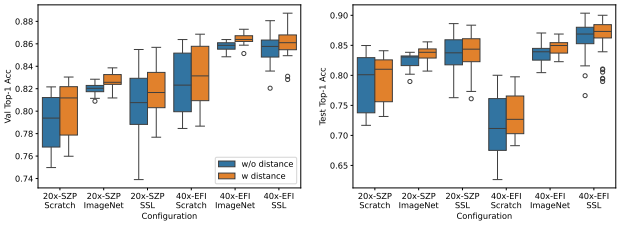
(a) (b) (c) (d) (e) (f)

Figure 5. Test set UMAP embeddings of ViT-B SimCLR (a, b), DINO (c, d) and MAE (e, f), using 40× EFI, with background. The legend in (f) also applies to (b) and (d).

The training results are displayed in Table A.7 and Figures 6, 7 and A.3. Like the baseline, models pretrained with SSL gain from additional distance information; however, as demonstrated in Figure 6b, the top-performing model regarding the test Top-1 Acc was actually trained without this supplementary information. The self-supervised learning methodologies examined in this study lack mechanisms to integrate additional distance information during the pretraining phase. This constrains the potential benefits that distance information could confer within this context.

As illustrated in Figure A.3 (and Figure 7), Vision Transformers tend to derive greater advantage from self-supervised pretraining relative to ResNet architectures.

Increasing model size does not always lead to better performance, given that the best overall model considering test



(a) (b)

Figure 6. Influence of distance information on val (a) and test (b) Top-1 Acc for the scan settings and pretraining types. SSL represents the combined results of SimCLR, DINO and MAE. 480 models trained with background only in 96 distinct configurations. The legend of (a) also applies to (b).

Top-1 Acc is a DINO ViT-T and not a ViT-B (see Figure 7d). However, ViT-Bs tend to be the best models regarding the validation metrics and are also often within the top models when it comes to the metrics on the test set. DINO ViT-B models for example exhibit comparable or better performance in test Adjacent-1, test Adjacent-1 Dynamic and validation accuracy measures when compared to DINO ViT-T models, suggesting that worse test Top-1 Accuracies are primarily due to misclassifications occurring in borderline cases. We therefore hypothesise that increasing the model size while also pretraining on large amounts of data is beneficial in most cases, which coincides with prior work on self-supervised learning [5, 6, 14].

For the 20× SZP setting MAE generally performs best as it scores highest in most validation and test metrics (see Figures 7a and 7d). We hypothesise that because numerous cells in the 20× SZP setting are already blurred and out of focus, applying additional augmentations - integral to the efficacy of methods such as SimCLR and DINO - may prove detrimental on this data, resulting in suboptimal priors. This effect can be observed even when compared to a pretrained ImageNet baseline from an out-of-domain context, which generally outperforms SimCLR and DINO in the 20× SZP setting with ResNet backbones. Reconstruction-based methods on the other hand require less aggressive data augmentation strategies and are therefore more adept at preserving detailed textural features. The increased Adjacent-1 Dynamic Acc of SimCLR and DINO are likely related to textural details being less important than the general shape of the cell for this specific metric.

For the higher quality 40× EFI images the trend of MAE outperforming SimCLR and DINO changes, especially in the case of test accuracies (see rightmost part of Figures 7a to 7f). Here MAE leads to worse results than SimCLR and DINO in Top-1 Acc and Adjacent-1 Dynamic Acc. While MAE's validation accuracies do improve in the 40× EFI setting, the test Top-1 Acc and Adjacent-1 Dynamic Acc decrease compared to the 20× SZP setting. Based on the

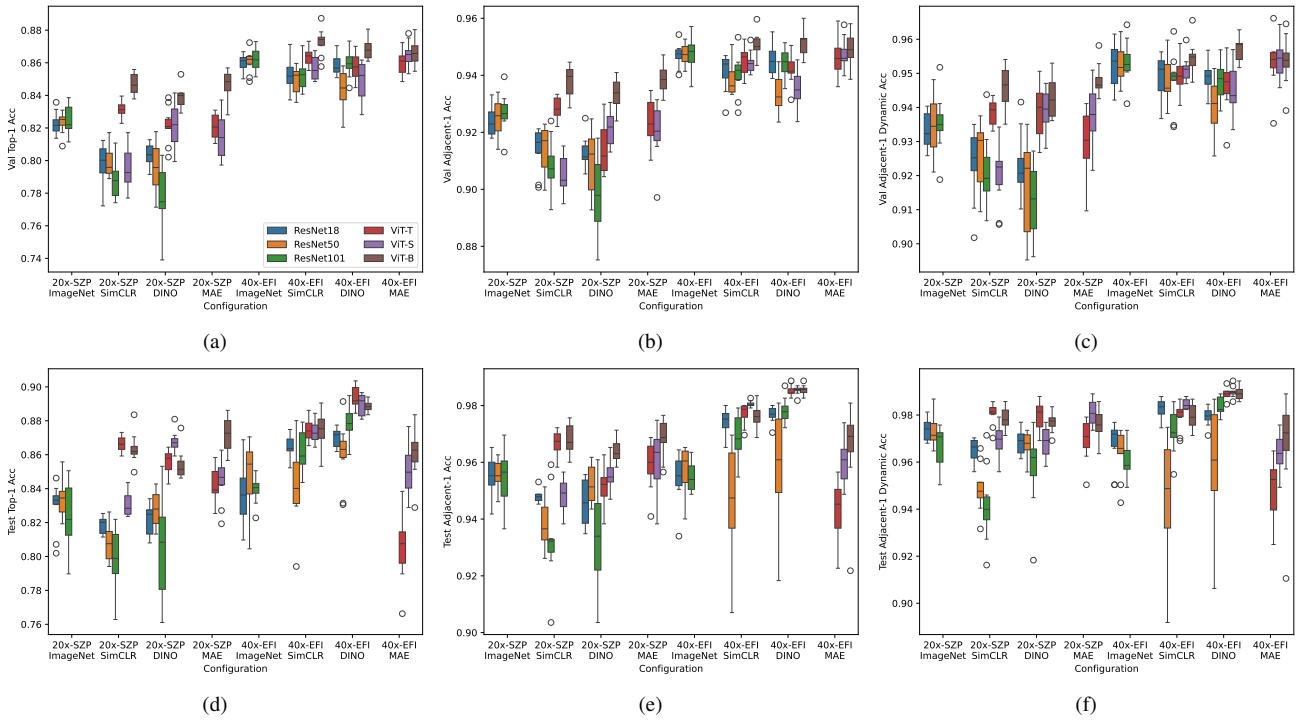


Figure 7. Boxplots of validation (a-c) and test (d-f) Top-1, Adjacent-1 and Adjacent-1 Dynamic Acc of all cross-validation models for every configuration with background. 360 models trained with background only in 72 configurations. Legend of (a) also applies to (b)-(f).

assumption that more information, through higher magnifications and cellular details, enhances performance, we plan to explore this finding further in future research. DINO generally performs best on the $40\times$ EFI test and very similar to the other self-supervised methods on $40\times$ EFI validation sets. For these higher quality images, SimCLR's and DINO's augmentation-heavy approach becomes an advantage instead of a disadvantage, as they both generalise well to the test set. In the $40\times$ EFI setting with ResNet backbones, both SimCLR and DINO surpass the ImageNet baselines on the test set, which previously held an advantage in the $20\times$ SZP setting.

5. Conclusion

In this paper, we examined the problem of classifying cells within a dynamic developmental process in static WSIs. We conducted various analyses, starting from the creation of the underlying data, which allowed us to gain several new insights which were progressively utilised for the training of neural networks. We performed an analysis of inter- and intra-annotation variability, quantitatively studied the impact of scan quality on the performance of the classification networks and extensively evaluated different deep learning techniques and their impact on task performance. Inter- and intra-variability can be accounted for when classifying into potentially ambiguous cell stage labels, e.g. by metrics such as the Adjacency-1 Accuracy. The deep learning models generally performed better with as much (context) information as possible, whether from sharper textures of higher quality images, background, or distance information. We showed that the data quality has a high impact on classification performance, that can be partially alleviated by using reconstruction based MAE as pretraining. When the quality and resolution of the data is high, however, the augmentation based approaches SimCLR and DINO yield better results. The results also show that Vision Transformers and larger model sizes are in most cases more suitable for the self-supervision techniques evaluated on our data. Overall our analyses provide new insights into the training of classification networks for dynamically developing cells, but also into aspects that influence the performance of different SSL techniques on histological image data of varying qualities.

Acknowledgements

ST and BR would like to thank the Human Frontier Science Program [RGP0057/2021]. JK, SK, JG and BR would like to thank the Deutsche Forschungsgemeinschaft (DFG) - CRU326. The calculations for this work were performed on the HPC cluster PALMA II of the University of Münster with special thanks to Sebastian Potthoff.

References

[1] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023. 3, 4

[2] Stuart Baum. The pas reaction for staining cell walls. *Cold Spring Harbor Protocols*, 2008(8):pdb–prot4956, 2008. 3

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 3

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. 4

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3, 4, 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4, 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[8] Shujian Deng, Xin Zhang, Wen Yan, Eric I-Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. Deep learning in digital pathology image analysis: a survey. *Frontiers of medicine*, 14:470–487, 2020. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[11] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. *International Conference on Machine Learning*, 2024. 3

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. 4

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3, 4, 5, 6, 7

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 4

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[17] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 4

[18] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023. 3, 4

[19] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7(1):29, 2016. 3

[20] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3344–3354, June 2023. 3, 4

[21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s), sep 2022. 6

[22] Jacqueline Kockwelp, Sebastian Thiele, Jannis Bartsch, Lars Haalck, Jörg Gromoll, Stefan Schlatt, Rita Exeler, Annalen Bleckmann, Georg Lenz, Sebastian Wolf, et al. Deep learning predicts therapy-relevant genetics in acute myeloid

leukemia from pappenheim-stained bone marrow smears. *Blood advances*, 8(1):70–79, 2024. 1, 3

[23] Jacqueline Kockwelp, Sebastian Thiele, Pascal Kockwelp, Jannis Bartsch, Christoph Schliemann, Linus Angenendt, and Benjamin Risse. Cell selection-based data reduction pipeline for whole slide image analysis of acute myeloid leukemia. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1825–1834, 2022. 1, 3

[24] Timo Kohlberger, Yun Liu, Melissa Moran, Po-Hsuan Cameron Chen, Trissia Brown, Jason D Hipp, Craig H Mermel, and Martin C Stumpe. Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. *Journal of pathology informatics*, 10(1):39, 2019. 3

[25] Melissa Linkert, Curtis T Rueden, Chris Allan, Jean-Marie Burel, Will Moore, Andrew Patterson, Brian Loranger, Josh Moore, Carlos Neves, Donald Macdonald, Aleksandra Tarkowska, Caitlin Sticco, Emma Hill, Mike Rossner, Kevin W Eliceiri, and Jason R Swedlow. Metadata matters: access to image data in the real world. *J Cell Biol*, 189(5):777–782, May 2010. 3

[26] Yun Liu, Timo Kohlberger, Mohammad Norouzi, George E Dahl, Jenny L Smith, Arash Mohtashamian, Niels Olson, Lily H Peng, Jason D Hipp, and Martin C Stumpe. Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. *Archives of pathology & laboratory medicine*, 143(7):859–868, 2019. 3

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[28] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, Feb. 2018. 6

[29] Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature methods*, 19(12):1634–1641, 2022. 3

[30] Liron Pantanowitz. Digital images and the future of digital pathology. *Journal of pathology informatics*, 1, 2010. 3

[31] Daniel D Polsby and Robert D Popper. The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale L. & Pol'y Rev.*, 9:301, 1991. 4

[32] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 3

[33] Swati Sharma, Joachim Wistuba, Tim Pock, Stefan Schlatt, and Nina Neuhaus. Spermatogonial stem cells: updates from specification to clinical relevance. *Human reproduction update*, 25(3):275–297, 2019. 1

[34] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016. 2

[35] Tzu-Hsi Song, Victor Sanchez, Hesham EI Daly, and Nasir M. Rajpoot. Simultaneous cell detection and classification in bone marrow histology images. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1469–1476, 2019. 2

[36] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021. 1, 3

[37] Ching-Wei Wang, Sheng-Chuan Huang, Yu-Ching Lee, Yu-Jie Shen, Shwu-Ing Meng, and Jeff L Gaol. Deep learning for bone marrow cell detection and classification on whole-slide images. *Medical Image Analysis*, 75:102270, 2022. 1

[38] Ling Zhang, Le Lu, Isabella Nogues, Ronald M. Summers, Shaoxiong Liu, and Jianhua Yao. Deeppap: Deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1633–1643, 2017. 2