

Multiple Instance Learning for Digital Pathology: A Review on the State-of-the-Art, Limitations & Future Potential

Michael Gadermayr, Maximilian Tschuchnig
Salzburg University of Applied Sciences

June 10, 2022

Abstract

Digital whole slides images contain an enormous amount of information providing a strong motivation for the development of automated image analysis tools. Particularly deep neural networks show a high potential with respect to various tasks in the field of digital pathology. However, a limitation is given by the fact that typical deep learning algorithms require (manual) annotations in addition to the large amounts of image data, to enable effective training. Multiple instance learning exhibits a powerful tool for learning deep neural networks in a scenario without fully annotated data. These methods are particularly effective in this domain, due to the fact that labels for a complete whole slide image are often captured routinely, whereas labels for patches, regions or pixels are not. This potential already resulted in a considerable number of publications, with the majority published in the last three years. Besides the availability of data and a high motivation from the medical perspective, the availability of powerful graphics processing units exhibits an accelerator in this field. In this paper, we provide an overview of widely and effectively used concepts of used deep multiple instance learning approaches, recent advances and also critically discuss remaining challenges and future potential.

Keywords: Multiple Instance Learning, Digital Pathology, Histology, Attention, Deep Learning

1 Motivation

For a large range of pathologies, microscopic evaluation of biopsies is the gold standard in clinical diagnostics. Examples are smear tests, analysis of cancerous tissues during operations and postmortem histological testing. Due to an increasing prevalence in combination with a decrease in the number of pathologists [1, 2], automated assistance tools will be of major importance in the near future. Digitization of slides is a first step towards efficiency enhancement which enables digital storage, easy transmission and digital processing.

Automated digital whole slide scanners are capable of digitizing complete microscopic slides within few seconds to several minutes [3]. In an iterative process, a plurality of neighboring patches are captured with a digital camera on a high magnification level. Finally, the patches are stitched in a way that the patch borders are indistinguishable and a single image can be accessed conveniently. Dedicated file formats [4] in combination with special image viewers allow quick and responsive visualization and interaction, in spite of the enormous size of the data. Hardware requirements for image visualization are nowadays also modest.

Digitization alone, however, is not capable of disrupting or clearly facilitating pathologists' daily routine [5], since digital workflows do not strongly differ from the conventional analog workflow. In both, digital and analog workflows, effective visual examination in clinical routine requires multiple processing stages on different resolutions. Typically a screening is performed first on a low resolution, followed by a detailed view on identified relevant regions of interest. And independent of whether slides are examined in the digital or analogue workflow, the enormous amounts of information in combination with time-pressure in clinical routine, exhibits a potential for missing relevant information within the data [6, 7, 8].

1.1 Potential & Limitations of Automated Image Analysis

The automated analysis of digital whole slide images (WSIs) exhibits a high potential for a plurality of applications in the field of pathology [5]. In recent literature, particularly segmentation [9, 10, 11] and classification [12, 13, 14] tasks were considered. State-of-the-art methods of resolution mainly consist of deep learning models. For both, segmentation and classification, convolutional neural networks are employed as the de facto standard method. Segmentation can be applied as a pre-processing technique, identifying either the shape, the area and/or the number of relevant regions of interest. The output segmentation maps can be either automatically processed or used to simplify the clinician's workflow by visualizing determined relevant regions. Classification approaches are typically employed for means of disease type or subtype categorization. While classification pipelines are mostly black-boxes, segmentation output can be easily interpreted and visually validated. For example, while a pathologist can easily assess the quality of a segmentation map with the naked eye, it is hard or even impossible to determine the reason (or a confidence) for a certain classification output. The categorical label also makes the integration into clinical workflows difficult. Even though probabilistic models [15] enable the computation of confidences in addition to categorical labels, a high level of transparency is thereby not achievable. Explainable deep learning techniques can be helpful, but straight forward approaches are not directly applicable due to the large size of whole slide images (WSI) [16]. A clear limitation of segmentation algorithms is given by the fact that neural networks typically require (manually) annotated training data provided in form of a segmentation map. This, in combination with the need for sufficiently large training data, represents a clear burden for training these models. Additionally, classification algorithms

are limited by the size of the neural networks’ input images. Even with high performance computers and graphic processing units (GPUs), gigapixel WSIs cannot be processed holistically (as a whole) using modern deep convolutional neural networks.

1.2 Motivation on Multiple Instance Learning

Multiple instance learning [17, 13] (MIL) exhibits a category of methods partly relaxing the limitations of both, segmentation and classification. Compared to conventional classification, MIL can be applied to whole slide images independently of the overall images size. Compared to the training of segmentation algorithms, there is no need to collect any local label information, e.g. by means of manually segmenting the regions of interest. Even though the ground-truth labels are available on whole slide image level only, multiple-instance learning algorithms are partly capable of generating local predictions during the inference phase. In that sense, MIL can be interpreted as an intermediate approach between segmentation and classification. The annotations, which are typically available on WSI-level only (there is one label per WSI), can be interpreted as weak labels.

1.3 Statistics

There is a clear trend towards MIL in the field of digital pathology. A Pubmed search based on the search-string¹

”multiple instance learning” AND (”digital pathology” OR ”histology”
OR ”histopathology” OR ”computational pathology” OR ”whole slide imag*”)

delivered 120 results, with 29 % (35) in 2021, 51 % (61) between 2020 and 2021 and 63 % between 2019 and 2021 as shown in Fig. 1. Besides established concepts, the focus of this work is on publications since 2020 showing novel technical approaches.

1.4 Contribution

In this paper, we provide an unstructured literature analysis of the basic building blocks of state-of-the art MIL. In addition to the basic MIL principles, we particularly focus on technical achievements developed, discussed and evaluated in recent literature since 2020 (motivated by the steep increase as shown in Fig. 1). Based on these approaches, we provide a structured and unified mathematical and textual description and a summary of similar techniques. Finally, we provide a critical discussion about technical opportunities and limitations of current approaches, with respect to the practical application in computational pathology and hardware requirements.

¹Link to Pubmed search

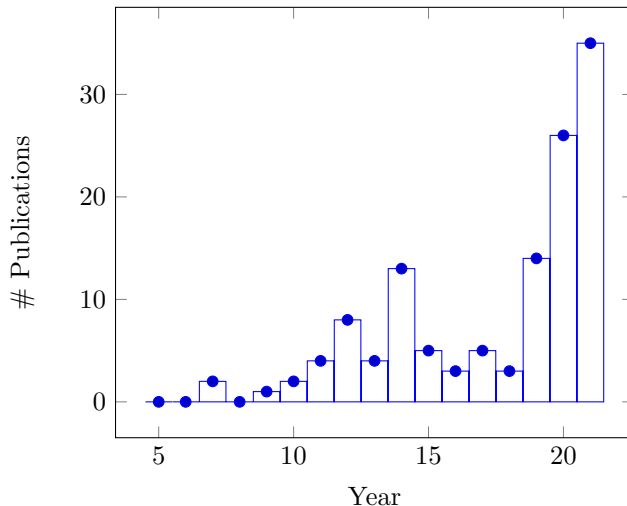


Figure 1: Pubmed search on the combination of *multiple instance learning* and *digital pathology*, as well as digital pathology synonyms (date: 2022-05-18).

The remaining part of this paper is structured as follows. In Section 2, the basic principals and a generic pipeline are presented. In Section 3, the state-of-the-art deep learning architecture and special components are introduced. In Section 4, focus is on the patch and feature extraction from the WSIs, exhibiting the first part of the generic pipeline. A critical discussion is provided in Section 5. Section 6 concludes this paper. We decided for a structure according to technical innovations rather than according to the publications. For that reason, individual papers are potentially mentioned in several sections.

2 Multiple Instance Learning in Pathology

Before the era of deep learning and deep neural networks, machine learning algorithms mostly consisted of two stages, the feature extraction stage and the classification stage. While the optimization of classification models has been employed by generic algorithms, feature extraction was often hand-crafted to the specific application scenario.

The era of deep learning changed this pipeline since deep convolutional neural networks enabled so-called end-to-end optimization of models with e.g. images as input and labels or label maps as output [18, 9]. The feature extraction stage, consisting of the concrete image filters, can thereby be automatically trained within the model. For some applications, however, it can be advantageous to separate a trained convolutional neural network into the convolutional part representing the feature extraction stage and the classification part of the model [19]. This enables, e.g. the combination of this new method of feature

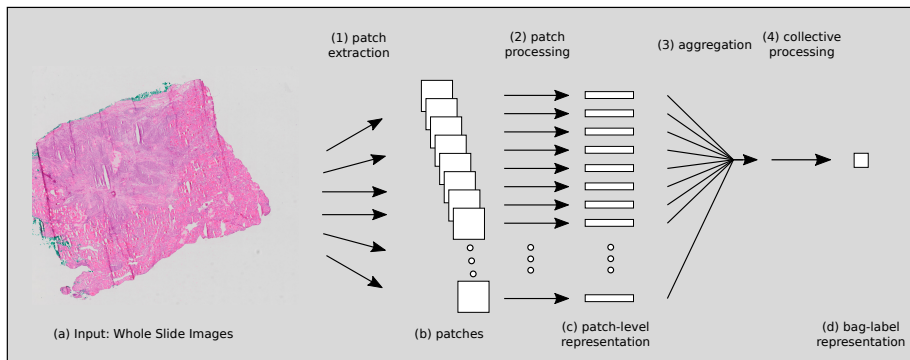


Figure 2: High-level perspective on MIL applied to WSIs. From the input images (a), patches (b) are extracted, followed by patch-level processing resulting in patch-level representations (c) and aggregation (several patch-level features to a single bag-level feature) and collective processing resulting in bag-level representations (d).

extraction with established classification models. Feature extraction models which were trained on a huge amount of data can thereby be combined with efficient classification models (with fewer parameters, such as support vector machines [11]) to achieve effective generalization in case that small training data is available only for the target application. Due to the often small number of available WSIs, this is particularly relevant for MIL.

On a very high level, MIL approaches in digital pathology can be abstracted to the definition outlined in Fig. 2. After extracting patches (1) from the original WSIs, each patch is first individually processed (2), followed by an aggregation (3) and a collective processing stage (4) which finally outputs a label corresponding to a "bag" of patches, which here corresponds to a WSI. This very generic pipeline can be slightly substantiated by restricting the type of data after the patch processing stage (2). In case that this stage finally outputs a scalar (e.g. scalar between 0 and 1) for each patch, the method is referred to as **instance-based MIL** approach. In case that this stage finally outputs a feature vector for each patch, the method is referred to as **embedding-based MIL** approach [13].

This minor differentiation has strong impact on the potential of the algorithms. Instance-based methods are capable of providing a final decision, individually for each patch. This output can be used to generate segmentation maps for complete WSIs indicating the relevance of different regions of interest. This advantage directly corresponds to the disadvantage of instance-based MIL. The restriction that the information of a patch must be represented as a single scalar value potentially limits the model's power. For that reason, embedding-based MIL approaches are typically more powerful in case that classification on whole slide image level is considered as final goal [20, 21].

The outlined pipeline remains the same, independent whether conventional

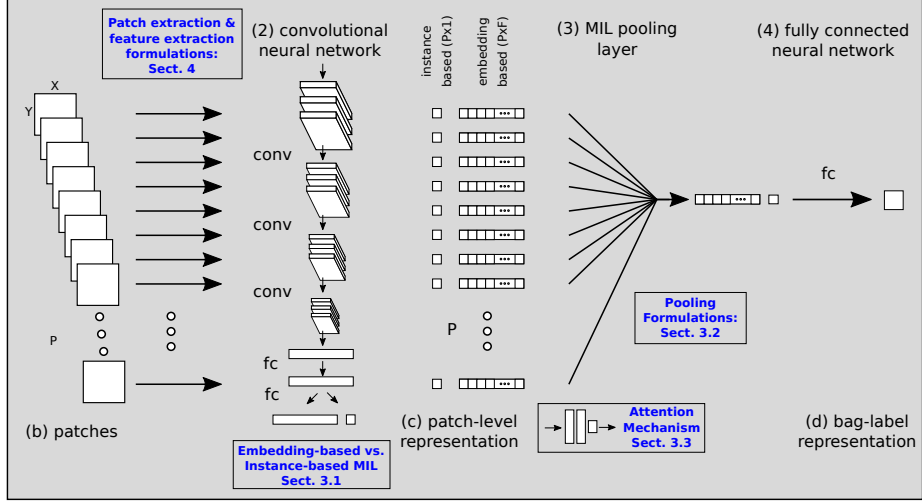


Figure 3: The scheme shown in Fig. 2, refined to show the outline of typical deep learning approaches and particularly the differences between instance-based and embedding-based MIL.

or deep learning-based models are employed. In the following, focus is on state-of-the-art deep learning architectures.

3 Deep Learning Architectures

The generic pipeline depicted in Fig. 2 can be implemented by a deep convolutional neural network in a quite easy way, which is also shown in Fig. 3.

As neural network input, we consider three dimensional samples of size $P \times X \times Y$, where the constants X and Y refer to the patch dimension (of the extracted patches) and P refers to the number of extracted patches. While X and Y must be chosen to fit the characteristics of the convolutional network (described in the next paragraph), in theory, P can be chosen freely. In Sect. 5, we discuss about further restrictions due to memory demands.

The first layers the patches are fed into, are convolutional layers (conv). Even though the input signal is three dimensional, only two dimensional filters are used here in a way that every single patch is processed individually (since the order in the third dimension is arbitrary and not meaningful). Often one of the well studied 2D ResNet models is used for that purpose [18]. As a final step, the output of the convolutional neural network is flattened for each patch individually, resulting in a matrix, e.g. with the different patches as (P) rows and a number of (F) features as columns.

This matrix is then aggregated by means of a pooling function. In theory, any differentiable function, projecting a $P \times F$ matrix (M) to a vector of length F is applicable here. Typical MIL Pooling functions are outlined in Sect. 3.2.

During this pooling operation, features per patch are converted into features per WSI. The final vector represents a descriptor for the complete histological slide.

To obtain a bag-level label in the end, either the output of the pooling function can be used or further neural network layers are applied. Typically for this purpose fully-connected (fc) layers are applied.

Since all operations are differentiable, this pipeline can be trained end-to-end, i.e. with pairs consisting of input patches and WSI labels. All parameters of the models can be trained at once using optimization algorithms, such as stochastic gradient descent with back propagation.

3.1 Deep Instance-based vs. Embedding-based MIL

Technically, with a minor configuration, this architecture can be either instance-based or embedding-based (Fig. 3 (c)). In case that the matrix M is of shape $P \times 1$ (exhibiting a column vector), a patch is represented by a single feature and the pipeline is referred to as instance-based. The single value automatically corresponds to a score or a confidence that the patch belongs to one of the classes. This is obtained when the convolutional neural network for patch-based feature extraction has exactly one output neuron per input patch. If M is a matrix of shape $P \times F$ with F greater than one, the approach is referred to as embedding-based.

3.2 MIL Pooling Function Formulations

We define the constants X and Y as the patch dimensions and P as the number of patches. The number of features is defined as F . The tuple

$$(\vec{v}_1, \vec{v}_2, \dots, \vec{v}_P)$$

contains for each individual patch, the corresponding feature vector. \vec{v}_p contains the features for the p -th patch represented as a column vector. \vec{v}_{pf} is the scalar value corresponding to the f -th feature of patch p .

Max-pooling is defined by \vec{y} , such that for each tuple element \vec{v}_p the maximum feature of all patches is computed by

$$y_f = \max_{p \in \{1, \dots, P\}} \vec{v}_{pf}, \quad f \in \{1, \dots, F\}.$$

Mean-pooling is defined such that for each feature, the arithmetic mean over the patches is computed by

$$y_f = \frac{1}{P} \cdot \sum_{p=1}^P \vec{v}_{pf}, \quad f \in \{1, \dots, F\}.$$

The soft-max function is a smooth approximation of the maximum function, defined as

$$y_f = r \cdot \log\left(\frac{1}{P} \cdot \sum_{p=1}^P r \cdot e^{\vec{v}_{pf}}\right), \quad f \in \{1, \dots, F\}$$

with r being an adjustable hyperparameter.

3.3 Attention Mechanism

Conventional pooling methods suffer from the limitation that the pooling method must be chosen manually and does not contain any trainable parameters. The attention mechanisms [22, 13] make use of the idea that each feature vector is weighted using the factor a_p , providing a measure for the importance of the patch with respect to the final decision. Attention based pooling is given by

$$y_f = \sum_{p=1}^P a_p \cdot \vec{v}_{pf}, \quad f \in \{1, \dots, F\}.$$

The parameters a_p are computed as follows [22, 13]

$$a_p = \frac{e^{w^T \cdot \tanh(W_1 \cdot v_p^T)}}{\sum_{i=1}^P e^{w^T \cdot \tanh(W_1 \cdot v_i^T)}}$$

with w being a trainable column vector of length F . The matrices W_1 ($F \times P$) contains trainable parameters. Since the equation for a_p uses the \tanh sigmoidal non-linearity, with almost linear behaviour in the interval $[-1, 1]$, a second non-linearity in the form of a sigmoid function (such as a logistic function) can be added (in combination with a second matrix W_2) to the calculation of a_p leading to the formulation

$$a_p = \frac{e^{w^T \cdot (\tanh(W_1 \cdot v_p^T) \odot \text{sigm}(W_2 \cdot v_p^T))}}{\sum_{i=1}^P e^{w^T \cdot (\tanh(W_1 \cdot v_i^T) \odot \text{sigm}(W_2 \cdot v_i^T))}}.$$

Rymarczyk et al. [23] adapt the so-called self-attention mechanism introduced by Zhang et al. [24] to model dependencies between instances within one bag in MIL. This is achieved by transforming the instances (referred to as a and b) into two feature spaces $W_3 \cdot \vec{v}_a$ and $W_4 \cdot \vec{v}_b$. The matrices W_3 and W_4 are (trainable) matrices of size $\frac{F}{k} \times F$ with k being a hyper-parameter to reduce the dimensionality. These feature spaces are combined using the inner product $s_{ab} = \langle W_3 \cdot \vec{v}_a, W_4 \cdot \vec{v}_b \rangle$ (to measure similarity) and are further employed to calculate β_{ab} such that

$$\beta_{ab} = \frac{e^{s_{ab}}}{\sum_{i=1}^P e^{s_{ab}}}.$$

Based on β_{ab} , for a certain instance (a), o_a can be calculated using the following expression, with W_5 (which is a $\frac{F}{k} \times F$ matrix) and W_6 ($F \times \frac{F}{k}$ matrix) as further trainable matrices such that

$$\vec{o}_a = W_6 \cdot \sum_{p=1}^P \beta_{ap} \cdot W_5 \cdot \vec{v}_p ,$$

with the sum (\sum) being an element-wise addition. This leads to a mapping of the original tuple $(\vec{v}_1, \vec{v}_2, \dots, \vec{v}_P)$ to the transformed space $(\vec{w}_1, \vec{w}_2, \dots, \vec{w}_P)$ defined by

$$\vec{w}_p = \mu \cdot \vec{o}_p + \vec{v}_p$$

using a trainable scaling parameter μ . Finally, this transformed space can be used to obtain the pooling, by utilizing the attention based pooling method, leading to

$$y_f = \sum_{p=1}^P a_p \cdot \vec{w}_{pf}, \quad f \in \{1, \dots, F\} .$$

Shao et al. [25] proposed another variation, which also makes use of a transformer based self-attention. This self-attention is used to encode the interactions of sequence tokens and to add positional information. This again leads to a transformed space which can further used by a final pooling method.

Oner et al. [26] developed an approach based on so-called distribution pooling. Based on the Gaussian assumption, the parameters of the normal distribution are estimated, rather than single scores, such as mean or max. Here the parameters of the marginal distributions are estimated, individually for each feature (since capturing the joint distribution is computationally infeasible [27]).

3.4 Instance and Embedding based MIL Combinations

Embedding-based MIL is typically more powerful, at least as it comes to the classification of complete WSIs [20], but does not enable a scoring for single patches (finally leading to confidence maps). Li et al. [28] proposed a combination of instance-based and embedding-based MIL. The proposed dual stream MIL approach jointly learns an instance-based and a embedding-based classifier, using a dual-stream architecture. One stream uses a standard instance-based approach combined with max-pooling as show in Sect. 3.1 to identify the highest scoring instance. The second stream computes an attention score for each instance by measuring its distance to the so-called critical instance which is the patch showing the maximum score. To obtain a final decision on bag-level, both scores are averaged. The attention weight in this approach is computed in a efficient way by computing the inner product between the critical instance and each individual instance. The method consisting of instance-based and bag-based MIL aims to combine the advantages of both approaches. In addition, the authors make use of the self attention mechanism (see Sect. 3.3).

3.5 Confidence Map Generation

In case of instance-based MIL, the instance scores can be directly used as local confidence measures. Li et al. [28] proposed a hybrid approach based on a combination of both, instance and embedding-based MIL combining the advantages of both techniques. Oner et al. [26] suggested a pure embedding-based MIL approach, which delivers final decisions or confidences only on bag (WSI) level. To obtain maps, first a bag of samples in a WSI within a certain region is collected and then processed similarly to complete WSIs. The predicted value for the bag is finally assigned to the center of the region. This step is repeated for each point in a rastered grid, leading to a variably resolved confidence map.

3.6 Clustering

Sharma et al [29] proposed a pipeline which makes use of local clustering as a special type of patch selection (but not for aggregation). Clustering can be employed to identify instances (patches) showing similar image features corresponding to similar image information. By performing clustering (here k-means) individually for each WSI and selecting a fixed number of instances for each cluster, the authors hypothesize that the relevant information of a WSI is more accurately approximated. Besides the clustering approach, the authors proposed an end-to-end pipeline making use of the attention mechanism and including a loss composed of the bag loss, the instance loss (similarly to 3.4) as well as a loss based on the Kullback-Leibler divergence [29] between patches. The latter is applied to regularize the high instance variance of attention distribution observed in similar positive instances.

3.7 Proxy Labels

Lerousseau et al. [14] introduced proxy labels to the patches of the WSIs, depending on their instance based scores. Patches of the positive class are labeled as

- 1 such that $\alpha\%$ of the patches with the highest probability are of class 1 and
- 0 such that $\beta\%$ of the patches with the lowest probability are of class 0.
- Other patches are discarded from the loss computation.

The parameters α and β are optimized during exhaustive search. These proxy labels can be interpreted as a quantisation of the patch level probabilities. This quantisation (or classification) enables a binary segmentation of the WSIs without requiring to choose an additional threshold value.

3.8 Siamese MIL

Yao et al. [30] proposed a MIL approach making use of a Siamese architecture and attention-based pooling. Focus here is on survival prediction which slightly

differs from most other classification tasks. Dedicated to ordinal scaled samples, this approach makes use of two parameter sharing fully-convolutional networks processing samples of two different WSIs. The parallel stage is followed by the aggregation stage. The loss is computed based on the outcome of two individual images. The goal is that, for pairs of subjects, the ones with the higher risk (lower survival rate) are ranked higher. The loss function contributes to the overall concordance by penalizing any discordance in any values of higher risk patients if they are greater than those of lower risk patients. Since the labels are not absolute, but allow a relation between two subjects, all other approaches for classification cannot be used for that use case.

3.9 Domain Adversarial Learning

Hasimoto et al. [31] proposed a technique based on domain adversarial training in order to optimize the feature extraction stage in the sense that the features are invariant to variations in stain intensity. In parallel to the final layers (attention-based pooling (3) and fully connected neural network (4) in Fig. 3), a discriminator, consisting of a simple neural network, is trained to predict the domain. In this work, each patient is treated as an individual domain so that no additional knowledge (label) on the staining condition is needed.

3.10 Confidence Scoring

Ianni et al. [19] proposed a simple yet effective method for providing a confidence score on WSI (bag) level. For that purpose, prediction for each individual WSI is performed multiple times. For each run, a random subset of the neurons of the neural network is omitted (here 70 %) leading to a distribution of labels rather than a single discrete label. The distribution of predictions can be transformed into a confidence score.

4 Patch & Feature Extraction

Here we focus on the first stage of the generic pipeline, consisting of patch extraction and feature extraction.

4.1 Patch Extraction

Patch extraction is typically done randomly or in a rectangular grid for both training and testing. If a confidence map should be generated, sampling for the inference phase need to be performed in a regular grid. To increase resolution, patches can be extracted with overlaps. The patch size varies between 150×150 and 224×224 pixels [20, 14] showing typical input sizes of two dimensional convolutional networks [18]. Patches showing background only or marginal tissue only are mostly omitted by means of handcrafted methods.

Since there is a correspondence between patch size and memory consumption and the number of patches and the memory consumption, both variables need to

be selected with the memory restrictions in consideration. Increasing the number of patches (P) leads to a linear increase in the memory needed to store the feature maps of the convolutional network (Fig. 3 (2)). Increasing the patch size (X, Y) requires a changed architecture (particularly the fully-connected layers in (2)) with varying impact on the memory consumption. However, particularly the first feature maps of large images typically show a large memory footprint due to the large size of the image data in combination with the number of feature maps.

Basic patch selection is typically performed as preprocessing step. Patches are extracted in regions showing mainly tissue and no background. Attention-based pooling has a similar effect, since invaluable information does not contribute to the final image representation. However, compared to a separate patch selection stage, an implicit selection in an end-to-end approach corresponds to an increased memory footprint. Ianni et al. [19] introduced patch selection based on a CNN pipeline including multiple stages, consisting of image normalization, patch selection and classification. For patch selection, the authors make use of a supervised setting by collecting segmentation masks from medical experts. The setting thereby slightly differs compared to the typical MIL setting with end-to-end training. The advantage of this approach is given by the fact that for training the MIL approach, only relevant data is used instead of randomly selected data.

4.2 Feature Extraction

For feature extraction, typically convolutional neural network architectures are utilized. When applying these models for MIL, there exist several options. The networks can be trained from scratch without any pre-training. Pretraining based on large datasets (such as the ImageNet dataset [32]) can be used to appropriately initialize the weights. These weights can be directly used (pretraining-only) or be further optimized on the specific training data set (pretrained CNN). Finally, unsupervised or self-supervised approaches can be applied.

4.2.1 Pretrained-Only CNN

A very generally applicable approach for feature extraction is the employment of pretrained networks (as done by Lerousseau et al. [14]). Training a powerful CNN, such as a ResNet architecture [18] with a large data set, e.g. the ImageNet dataset, is a well studied approach in the field of image analysis [33, 34] and the field of digital pathology [11]. In the latter case, training from scratch is often inhibited since patch-level annotations are not available. Typically, the output of such a CNNs last convolutional layer is flattened and used as generic yet powerful image representation. The advantage of this setting is clearly the generic and efficient applicability. Training could be inhibited by the absence of (large) labeled training data, or by the absence of computing infrastructure. Pretrained CNN models can also be downloaded and immediately used. The disadvantage of pretrained CNNs is given by the fact that they are not optimized with respect

to the application scenario. Even though generic features proved to work well in many domains, individual training can lead to improved representations, in turn leading to higher scores in the end.

4.2.2 Transfer Learning

To exploit both, powerful large image data sets on the one hand and the incorporation of peculiarities of the specific application scenario on the other hand, pretrained networks can be adapted as performed in the case of Zhao et al. [35]). A CNN, pretrained on a large data set, is adapted to the specific data set, by initializing the weights accordingly. Thereby the need for huge amounts of data can be relaxed since the whole feature extraction part is initialized appropriately. Even though the parameters are well initialized, in case of small data sets, it is important to keep an eye on overfitting [33].

4.2.3 Autoencoders

Autoencoders are neural networks which are optimized to generate a latent code, based on the criterion that the output regenerates the input of the network. Since the latent code typically shows lower dimensionality, the compression during encoding is intended to maintain the important characteristics only, providing a good image representation. Variational autoencoders [36] adapt this idea by making use of multivariate latent distributions instead of basic feature vectors.

For WSI classification, Zhao et al. [35] employed a variational autoencoder, combined with a generative adversarial model (VAE-GAN [37]) to emphasize on generating a latent representation while enabling realistic reconstructions of the input images. To obtain a rather small, yet discriminative subset of features for each patch, the authors additionally perform feature selection to generate a compact description. To identify redundant features, maximum mean discrepancy was applied. The authors proposed an approach based on a graph-CNN to obtain decisions on a bag level, which is different from most other approaches.

4.2.4 Contrastive Learning

Li et al. make use of contrastive, self-supervised learning for feature extraction [28]. Specifically, SimCLR [38] is deployed for learning representations for individual patches. This approach makes use of a contrastive learning strategy by training a CNN to associate the subimages from the same WSI in a set of patches. The model is trained to maximize the agreement between the patches from the same WSI using a contrastive loss. After CNN training, the feature extractor is used to compute the representations of the training samples for downstream tasks of embedding MIL applications.

4.2.5 Multi-Scale Features

Li et al. [28] proposed a pyramidal fusion mechanism to obtain multi-scale WSI features. To aggregate descriptive information from several resolutions, in this approach, patch-level descriptors from two resolutions are combined by means of vector concatenation. In addition, Tschuchnig et al. [39] investigate the effect of three different approaches for combining several resolution. Besides the vector concatenation approach [28], they also consider the concatenation of histograms (after the bag-of-words aggregation) and the aggregation of features from all scales into one single histogram. This approach is not based on end-to-end deep learning, but uses bag-of-words clustering and a support vector classifier.

Hashimoto et al. [31] suggest to first train individual feature extractors on single resolutions (see also Sect. 3.9). The individually trained feature extractors are finally combined based on an attention-pooling approach.

5 Discussion

Recently, a large number of approaches for classifying WSIs by means of MIL have been developed as outlined in Sect. 3 and 4.

Despite of the fact that a plurality of different medical application scenarios have been investigated in the publications, from a high level perspective the tasks are similar for most approaches. We identified the following goals which can be reached with MIL: The two obvious goals refer to the classification of patches (instances) as well as WSIs (bags). While the classification of patches requires an architecture which aggregates the patch information to a single value (instance-based or a combination 3.4), the classification of WSIs can be performed with any approach. Although the classification of WSIs is an obvious goal, the hard label (numeric or even binary value) thereby obtained is not optimal to be included into a clinical workflow. It could be used in a setting where the computer generates the decision or a computer confirms the decision of a pathologist. While the second approach is more obvious (to be applied in the near future), it is not suited to increase efficiency of the clinicians' workflow. Most models also do not provide any confidence for a decision on the level of WSIs. Patch-based decisions are easier to interpret and can be integrated in a clinical workflow, for example to provide an estimate of relevant regions-of-interest to a pathologist. Special architectures even allow a more precise segmentation (beyond the resolution of patch sizes only) [14]. This allows the provision of a smooth heat map indicating relevant regions-of-interest. For integration into an effective and accepted software tool, in our opinion this approach has a high potential. Confidences on the WSI-level exhibit a further feature for interpreting the certainty of a neural network (see Sect. 3.10).

5.1 Limited Hardware Ressources

In spite of the ongoing development and improvement of hardware and particularly graphic processing units, memory is still a limiting factor for processing

digital WSIs. An image with a size of one gigapixels clearly fits into the memory of a single GPU. However, the first convolutional layer of a fictitious huge CNN processing such an image immediately breaks the limits of most current GPUs during training. This provides a strong motivation for the development of more efficient MIL approaches. By using a subset of the available data, arranged into patches, the memory consumption can be adjusted. Even if hardware could cope with the huge images as a whole in future, it is still highly questionable if deep learning architectures could be effectively trained at any point in time. The huge images would also require deeper architecture to aggregate the data and to focus on the relevant features. Since also the number of available WSIs is often limited, we expect that training such architectures would fail in most cases, independently of hardware limitations.

Although MIL is capable of relaxing the challenge regarding hardware limitations, the problem is thereby not solved. It is noteworthy to mention that MIL approaches do not analyse the whole WSI, but an often randomly sampled subset only. Typically the input dimension is in the sphere of 100^3 pixels where the first two dimensions refer to the image dimensions and the last refers to the number of patches extracted per WSI. Compared to a complete WSI (of a gigapixel), this corresponds to only one percent of the overall image data. Since large areas of the image are white, the patch-wise processing eliminates large unneeded information. However, in case of large tissue probes, clearly not all areas are processed and thereby considered. Certain approaches focus on this challenge and propose methods for effective patch selection or sampling [29]. Now developments in the field of GPUs will also enable a denser sampling.

We would like to further point out that there is a trade-off (with respect to memory consumption) regarding the size of the first two dimension (patch size) and the third dimension (number of involved patches). Increasing the patch size leads to a decrease in the number of patches and vice versa if the memory consumption should remain stable. We did not identify a publication explicitly focusing on the evaluation of the best trade-off. However, we also did not identify clear justifications for the chosen settings. For that reason we expect that a deep analysis could lead to further improvements.

5.2 Limited Training Data & Data Augmentation

The overall amount of training samples on WSI basis are mostly limited and in the area of tens to hundreds, while the amount of pixels and extractable patches per WSI is huge. Since during end-to-end training, many samples correspond to a single WSI, the lack of data exhibits a clear challenge. Interestingly, data augmentation is explicitly considered in few work only.

Gadermayr et al. [40] proposed a random patch sampling strategy for a conventional (not deep learning-based) approach. Based on a large amount of extracted patches per WSI, a subset is randomly selected and used for training and inference. Thereby, a large number of samples can be created for each individual WSI. The trade-off is given by the fact that reducing the amount of sampled patches reduces the quality of an individual representation, while

increasing the potential variability. This simple approach allows performing augmentation also during inference combined with aggregation, for example using majority voting.

Li et al. [41] proposed a data augmentation strategy for MIL-based CT image classification. They present a strategy to sample virtual sets of patches based on the knowledge obtained from the attention mechanism. By keeping the distribution of samples with a high attention and a low attention similar, new virtual bags are created by random sampling. Even though this method is not developed for digital pathology, it can be assumed that it is also applicable for WSIs.

Stegmüller et al. [42] proposed a technique for creating virtual patches out of existing patches, based on combining a pair of patches to one new patch. This does not solve the issue of a small number of WSIs, but can be combined with one of the approaches above.

A different approach with similar effect as data augmentation is stain normalization [43, 44]. Ianni et al. [19] included image normalization by means of a convolutional neural network, which is trained end-to-end together with the classification network. In this study, focus was explicitly on real world data from multiple sites showing clear variability. Yao et al. [30] also proposed a method including (adversarial) stain normalization in the classification pipeline. A standalone image normalization (or image optimization) is performed by Gadermayr et al. [40], by means of an unpaired generative adversarial network approach (cycle-GAN). In this method, the stain normalization is a pre-processing step not included in an end-to-end pipeline.

5.3 Model Pretraining

Even though there exists a large amount of publications and many different data set in the field of digital pathology, the aggregation of data sets or the use of specifically pretrained models is not observed so far. Regarding the feature extraction stage, it is common knowledge to pretrain convolutional neural networks on publicly available large generic data sets (such as the ImageNet data set). Such techniques are also applied in the field of digital pathology. However, we did not observe publications focusing on the utilization or combination of more similar data sets showing medical or histological image content.

Firstly, digital pathology clearly distinguishes from other natural image data. Secondly, in the field of digital pathology, focus is often on similar tissue, such as cancer. Whereas organs show clearly different structure, cancer tissue shows high similarity, independently of the organ of interest. Also low-level tissue entities such as nuclei show similarities between organs.

5.4 Alternatives to MIL

Beyond classical MIL algorithms, there exist other approaches focusing on the classification of WSIs based on the same setting (with bag-labels only). These methods could also be interpreted as MIL models.

An example is the method proposed by Hou et al. [12]. The method consists of a classification on patch level. The labels are aggregated into histograms and finally used for training a supervised model. Thereby the method can be interpreted as so-called count-based MIL technique.

A similar method [11] uses pretrained networks for patch-wise feature extraction and performs aggregation of these features by clustering and the bag-of-words approach to obtain a feature histogram. Based on the histograms, finally a supervised classification model (support vector machine) is trained.

Particularly the latter approach proved to work well with a small number of WSIs, since the feature extraction stage is based on pretrained features only. Both models use shallow classification models, such as support vector machines, which are also easy to train with few dozens of samples only.

6 Conclusion

Recently, numerous different approaches for classification of WSIs by means of MIL were developed. In this work, we summarized frequently used deep learning architectures and focus on cutting edge literature and the containing novel aspects. Although the overall architecture in the majority of approaches is similar, we identified interesting and powerful specific modifications. The proposed approaches are generally applicable to different histological fields and are typically not handcrafted to any specific image or data set features. As existing limitations, we identified the absence of sufficient training data, and the availability of graphics processing units with a sufficient amount of memory. In the latter case, ongoing developments can provide a further boost by enabling the incorporation of more information from the huge histological images. To more effectively deal with small data sets, we identified research needs in the field of data augmentation as well as the use of transfer learning, based on similar, large available data sets.

References

- [1] Stephen W Barthold, Alexander D Borowsky, Cory Brayton, Rod Bronson, Robert D Cardiff, Steven M Griffey, Tan A Ince, Alexander Yu Nikitin, John Sundberg, and V E Ted Valli. From whence will they come? a perspective on the acute shortage of pathologists in biomedical research. *Journal of Veterinary Diagnostic Investigation*, 19(4):455–456, 2007.
- [2] Victor Mudenda, Evans Malyangu, Shahin Sayed, and Kenneth Fleming. Addressing the shortage of pathologists in Africa: Creation of a MMed Programme in Pathology in Zambia. *African Journal of Laboratory Medicine*, 9(1):1–7, 2020.
- [3] Matthew G Hanna, Orly Ardon, Victor E Reuter, Sahussapont Joseph Sirintrapun, Christine England, David S Klimstra, and Meera R Hameed.

- Integrating digital pathology into clinical practice. *Mod. Pathol.*, 35(2):152–164, 2022.
- [4] Sébastien Besson, Roger Leigh, Melissa Linkert, Chris Allan, Jean-Marie Burel, Mark Carroll, David Gault, Riad Gozim, Simon Li, Dominik Lindner, and Others. Bringing open data to whole slide imaging. In *Proceedings of the European Congress on Digital Pathology*, pages 3–10, 2019.
 - [5] Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of Pathology Informatics*, 9, 2018.
 - [6] Hélène Dano, Serdar Altinay, Laurent Arnould, Noella Bletard, Cecile Colpaert, Franceska Dedeurwaerdere, Benjamin Dessauvague, Valérie Duwel, Giuseppe Floris, and Stephen Fox. Interobserver variability in upfront dichotomous histopathological assessment of ductal carcinoma in situ of the breast: the DCISion study. *Modern Pathology*, 33(3):354–366, 2020.
 - [7] Mieke R Van Bockstal, Martine Berlière, Francois P Duhoux, and Christine Galant. Interobserver variability in ductal carcinoma in situ of the breast. *American Journal on Clinical Pathology*, 154(5):596–609, 2020.
 - [8] Andrew T Turk, Sylvia L Asa, Zubair W Baloch, William C Faquin, Giovanni Fellegara, Ronald A Ghossein, Thomas J Giordano, Virginia A LiVolsi, Ricardo Lloyd, and Ozgur Mete. Interobserver variability in the histopathologic assessment of extrathyroidal extension of well differentiated thyroid carcinoma supports the new american joint committee on cancer eighth edition criteria for tumor staging. *Thyroid*, 29(5):619–624, 2019.
 - [9] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, and Katharina Seiwald. U-Net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16(1):67–70, 2019.
 - [10] Aïcha BenTaieb and Ghassan Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *Proceedings of the Conference on Medical Image Computing and Computer Assisted Interventions*, pages 460–468. Springer, 2016.
 - [11] Maximilian E Tschuchnig, Philipp Grubmüller, Lea M Stangassinger, Christina Kreutzer, Sébastien Couillard-Després, Gertie J Oostingh, Anton Hittmair, and Michael Gadermayr. Evaluation of Multi-Scale Multiple Instance Learning to Improve Thyroid Cancer Classification. In *Proceedings of the International Conference on Image Processing Theory, Tools, and Applications*, 2022.
 - [12] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016.

- [13] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. 2020.
- [14] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions*, 2020.
- [15] Yasha Zeinali and Brett A Story. Competitive probabilistic neural network. *Integrated Computer-Aided Engineering*, 24(2):105–118, 2017.
- [16] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.
- [17] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.
- [18] 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.
- [19] Julianna D. Ianni, Rajath E. Soans, Sivaramakrishnan Sankarapandian, Ramachandra Vikas Chamarthi, Devi Ayyagari, Thomas G. Olsen, Michael J. Bonham, Coleman C. Stavish, Kiran Motaparathi, Clay J. Cockrell, Theresa A. Feeser, and Jason B. Lee. Tailored for Real-World: A Whole Slide Image Classification System Validated on Uncurated Multi-Site Data Emulating the Prospective Pathology Workload. *Scientific Reports*, 10(1):3217, 2020.
- [20] Joshua Butke, Tatjana Frick, Florian Roghmann, Samir F El-Mashtoly, Klaus Gerwert, and Axel Mosig. End-to-end Multiple Instance Learning for Whole-Slide Cytopathology of Urothelial Carcinoma. In *In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 57–68, 2021.
- [21] Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. Key instance detection in multi-instance learning. In *Proceedings of the Asian Conference on Machine Learning*, pages 253–268, 2012.
- [22] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *In Proceedings of the International Conference on Machine Learning*, pages 2127–2136, 2018.

- [23] Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zielinski. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1721–1730, 2021.
- [24] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, pages 7354–7363, 2019.
- [25] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Others. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems (NIPS)*, 34, 2021.
- [26] Mustafa Umit Oner, Jianbin Chen, Egor Revkov, Anne James, Seow Ye Heng, Arife Neslihan Kaya, Jacob Josiah Santiago Alvarez, Angela Takano, Xin Min Cheng, Tony Kiat Hon Lim, and Others. Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study. *Patterns*, 3(2):100399, 2022.
- [27] Mustafa Umit Oner, Jared Marc Song Kye-Jet, Hwee Kuan Lee, and Wing-Kin Sung. Studying The Effect of MIL Pooling Filters on MIL Tasks. *arXiv preprint arXiv:2006.01561*, pages 1–16, 2020.
- [28] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021.
- [29] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A Moskaluk, Sana Syed, and Donald Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Proceedings of the International Conference on Medical Imaging with Deep Learning*, pages 682–698, 2021.
- [30] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.
- [31] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3852–3861, 2020.
- [32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of*

- the *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [33] Yiting Xie and David Richmond. Pre-training on grayscale imagenet improves medical image classification. In *In Proceedings of the European Conference on Computer Vision*, page 0, 2018.
 - [34] Maayan Frid-Adar, Avi Ben-Cohen, Rula Amer, and Hayit Greenspan. Improving the segmentation of anatomical structures in chest radiographs using u-net with an imagenet pre-trained encoder. In *Proceedings of the International MICCAI Workshop on Reconstruction and Analysis of Moving Body Organs*, pages 159–168. 2018.
 - [35] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, and Others. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4837–4846, 2020.
 - [36] Ruoqi Wei and Ausif Mahmood. Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey. *IEEE Access*, 9:4939–4956, 2020.
 - [37] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *In Proceedings of the International Conference on Machine Learning*, pages 1558–1566, 2016.
 - [38] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *In Proceedings of the International Conference on Machine Learning*, pages 1597–1607, 2020.
 - [39] Maximilian E. Tschuchnig, Gertie J. Oostingh, and Michael Gadermayr. Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential. *Patterns*, 1(6):100089, 2020.
 - [40] Michael Gadermayr, Maximilian Tschuchnig, Lea Maria Stangassinger, Christina Kreutzer, Sebastien Couillard-Despres, Gertie Janneke Oostingh, and Anton Hittmair. Frozen-to-paraffin: Categorization of histological frozen sections by the aid of paraffin sections and generative adversarial networks. In *Proceedings of the International MICCAI Workshop on Simulation and Synthesis in Medical Imaging*, pages 99–109, 2021.
 - [41] Zekun Li, Wei Zhao, Feng Shi, Lei Qi, Xingzhi Xie, Ying Wei, Zhongxiang Ding, Yang Gao, Shangjie Wu, Jun Liu, and Others. A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning. *Medical Image Analysis*, 69:101978, 2021.

- [42] Thomas Stegmüller, Antoine Spahr, Behzad Bozorgtabar, and Jean-Philippe Thiran. Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. *arXiv Prepr. arXiv2202.07570*, 2022.
- [43] M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *Proceedings of the International Symposium on Biomedical Imaging*, pages 953–956, 2019.
- [44] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *Proceedings of the International Symposium on Biomedical Imaging*, pages 1107–1110, 2009.