Adversarial Stain Transfer for Histopathology Image Analysis

Aïcha BenTaieb and Ghassan Hamarneh
Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Burnaby, Canada

***Abstract*—It is generally recognized that color information is central to the automatic and visual analysis of histopathology tissue slides. In practice, pathologists rely on color, which reflects the presence of specific tissue components, to establish a diagnosis. Similarly, automatic histopathology image analysis algorithms rely on color or intensity measures to extract tissue features. With the increasing access to digitized histopathology images, color variation and its implications have become a critical issue. These variations are the result of not only a variety of factors involved in the preparation of tissue slides but also in the digitization process itself. Consequently, different strategies have been proposed to alleviate stain-related tissue inconsistencies in automatic image analysis systems. Such techniques generally rely on collecting color statistics to perform color matching across images. In this work, we propose a different approach for stain normalization that we refer to as stain transfer. We design a discriminative image analysis model equipped with a stain normalization component that transfers stains across datasets. Our model comprises a generative network that learns dataset-specific staining properties and image-specific color transformations as well as a task-specific network (e.g. classifier or segmentation network). The model is trained end-to-end using a multi-objective cost function. We evaluate the proposed approach in the context of automatic histopathology image analysis on three datasets and two different analysis tasks: tissue segmentation and classification. The proposed method achieves superior results in terms of accuracy and quality of normalized images compared to various baselines.**

***Index Terms*—Histopathology, Stain Normalization, Generative Learning, Adversarial Training.**

I. Introduction

Historically, histopathology and cytopathology have been the main tools utilized in the diagnosis of cancer. In fact, with the exception of rare cases, the diagnosis of cancer is primarily confirmed by pathologists' visual analysis of the morphology of histological sections under a microscope. The process of examining histological sections involves the preparation of tissue biopsies using colored natural or chemical staining agents that selectively binds to naturally transparent tissue components (e.g., nuclei and cells) [1]. An example of one of the most commonly used staining agent, Hematoxylin and Eosin (H&E), is shown in Figure 1.

Given the chemical nature of the staining procedure, many variables can alter the visual appearance of a given tissue section [4]. For instance, tissue types, reactivity to staining agents from different manufacturers, tissues' thickness, concentration of staining agents, as well as room temperature during preparation, are some of the many factors that can cause variations in the appearance of stained tissues. Besides staining, variations in the tissues' appearance can further be introduced during digitization reflective of the type of optical microscope used. Figure 1 shows examples of such variations.

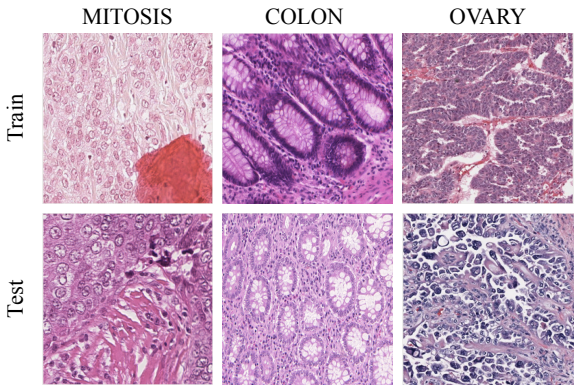

Fig. 1: Examples of staining inconsistencies. "MITOSIS": images from the ICPR 2014 challenge [2] belonging to the same class of nuclear atypia; "COLON": images from the MICCAI 2015 GlaS challenge [3] and represent benign colon adenocarcinomas. "OVARY": images from a dataset of whole slides images of subtypes of ovarian cancer and correspond to high-grade serous ovarian carcinomas.

As the tissues' visual appearance directly impacts the quality and accuracy of clinical diagnosis, controlling the amount of variation is paramount to a reliable diagnosis [5]. One of the existing solutions towards reducing variability relies on standardizing the protocols used for tissue preparation. However, these protocols currently involve many manual tasks (e.g., sectioning and applying staining agents), which implies that a complete standardization may be practically unattainable. In practice, quality-control methods involve subjective assessments of stain quality or inter-laboratory comparisons of stained tissue sections. A more practical solution towards normalizing stained tissue slides may involve automatic software-based systems for digital pathology. Such software-based stain normalization will be necessary for the complementary image interpretation systems (e.g., machine-learning based models for cancer classification or tissue segmentation), as they are sensitive to variations in tissue appearance, especially to color variations, as they generally rely on color-based features [6].

Generally, existing works proposed in the area of automatic image analysis address the problem of stain inconsistency by either (i) bypassing the problem using only grayscale images; (ii) relying on color-based data augmentation to synthesize new images in an attempt to enforce robustness to color inconsistency in the analysis model; or (iii) pre-processing images using color-matching strategies, usually called stain normalization techniques, in which all images from a given dataset are mapped to a user-selected reference (or template) image [7]–[15]. Methods that ignore color information (i.e. (i) above) or rely on data augmentation for learning color-invariant features [16] (i.e. (ii)) generally favor texture-based features [6] but have the major limitation of ignoring clinically relevant information captured through colors. As per stain normalization techniques used as pre-processing steps to normalize images (i.e. (iii)), we empirically demonstrate in this work their sensitivity to the choice of reference image.

We focus on the problem of stain inconsistency in the context of automatic histopathology image analysis and we aim at creating a model that would ultimately facilitate the applicability of automatic histopathology image analysis systems, such as nuclei segmentation or cancer classification systems, across pathology centers and regardless of the image acquisition procedure. Specifically, we propose a novel methodology for transferring stains across different datasets describing a similar pathology but with different staining appearance. We assume that such a model should couple stain normalization with the image analysis task without involving pre-processing images or re-training the analysis system. Also, we believe such system should not rely on a single given template reference image but should rather learn how best to transfer stains across datasets by leveraging datasets distributions. Based on these assumptions, we propose a fully trainable framework in which stain normalization is modelled as an adversarial game [17] and is performed jointly with a specific image analysis task. In a nutshell, given two datasets of stained images with different appearance, a generative model implicitly learns the color distribution of images in order to transfer stains across datasets and generate novel images with conserved texture properties but transformed color or stain appearance. In order to synthesize realistic transformed images, the generative model competes against a task-specific discriminative model in an adversarial framework [17]. We perform extensive experiments on 3 histopathology datasets including large-scale whole slide images for two image analysis tasks (i.e. image segmentation and classification).

II. Related Works

Previous works related to the proposed method broadly fall into three categories: (1) color-free and data augmentation strategies; (2) statistics-based color matching techniques; and (3) style transfer with generative modelling. Note that we only focus on existing strategies that handle staining inconsistencies in the context of automatic systems for histopathology images (e.g. these methods are proposed as part of an automatic systems' pipeline).

Color-Invariant Methods: Many works have shown the importance of texture for the analysis of histopathology images. In fact, texture features such as grayscale co-occurrence matrices, local binary patterns, wavelet transforms, among others, have been widely adopted in histopathology image analysis [18], [19]. In these works, images are converted to grayscale in order to bypass staining variations across datasets.

Another approach that indirectly results in learning features robust to color variation is color-based data augmentation. This strategy is especially relevant to recent deep learning models and consists in generating images by performing random or weighted color perturbations (via the so-called "fancy PCA" approach [20]) altering RGB intensities to generate novel images. This strategy is based on the assumption that objects' identity is invariant to changes in color intensity and illumination and has been used in many deep learning applications to histopathology resulting in learning more robust features.

In the context of histopathology images, color intensities have a special meaning as they relate to the concentration of stains and thus are descriptive of the biological composition of tissues. Using grayscale images results in mixing the staining concentrations and leads to images representing the total concentration of all tissue components instead of the relative concentration of each. Bypassing color implies omitting the wealth of information contained in tissue images and is, thus undesirable. The importance of color in pathology has been demonstrated in several studies; it has been confirmed that, despite being impacted by staining inconsistencies, clinicians learn to adapt to color variation when making their diagnosis [21]. A similar concern can be raised regarding color-based data augmentation strategies and their implication on the color-invariant predictions of deep learning models.

Color-matching methods: Among the works that do use color in the training of histopathology image analysis systems, many focus on including a pre-processing step in the automatic system's pipeline in order to normalize images from different datasets (e.g train and test sets in machine learning based frameworks). Existing stain normalization techniques consists of mapping the RGB colors of source images to a user-provided reference image. Note that such mapping can result in normalizing training images to a test image or vice versa.

There exist different strategies proposed to perform color-mappings. For instance, Reinhard et al. [10] proposed to match statistics of color histograms of a reference and target image after transformation of RGB images to the de-correlated Lab color space. Histogram matching techniques generally assume that the proportions of stained tissue components for each staining agent are similar across images being normalized. This assumption does not hold in most tissue images as the proportion of tissue components generally varies.

In order to overcome the limitation of histogram-matching, techniques based on stain separation prior to color normalization have been proposed [9], [15], [22], [23]. Images are first deconvolved into their principal staining constituents then normalization is performed on each staining channel separately, which involves a mapping between reference and source image. Different methods have been used to extract

the main staining constituents of a color image. Among the most popular ones, color deconvolution [18] decomposes an RGB image into its staining components by estimating a "staining matrix" that represents the RGB colors of each stain present in the tissue image. The staining matrix can be fixed or learned based on the statistics extracted from the dataset. Khan et al. [9] proposed to estimate the staining matrix using a supervised learning approach and pixel-level statistical color descriptors (e.g., color histogram). Other works for stain separation are based on performing non-negative matrix factorization [22], [23] to cluster an image into its principal components which, in the supervised form, requires knowledge of the number of clusters. Unsupervised forms of stain separation involve clustering the image using optimization algorithms such as Expectation-Maximization [12]. However, these stain separation techniques may not work optimally when there is an imbalanced representation of tissue clusters in the image, which may result in the algorithm mapping similar appearing regions from distinct tissue classes to the same cluster. To avoid these cases, extensions to these clustering techniques involved identifying distinct tissue clusters in a feature space and performing a cluster-to-cluster distribution alignment across reference and target images [15].

Besides all limitations related to the stain separation process, color-matching normalization techniques restrict the matching to be performed based on a single reference image, which, if not representative of the different object categories present in the image, may result in normalization errors.

Style Transfer and Generative Learning: In this work, we approach the problem of stain normalization as a style-transfer problem [24], [25], in which we aim at transferring the staining appearance of tissue images across different datasets. In computer vision, style transfer is generally seen as a problem of illumination and texture transfer where the goal is to synthesize a texture from a source image while preserving the semantic content of a target image. Style transfer involves finding image representations that independently model variations in the semantic image content and the style in which it is presented. Most recent works in style transfer have shown that deep feature representations from CNNs are powerful representations that can be used to independently manipulate the content and the style of natural images. Gatys et al. [24] propose to generate new images with a given reference "style" by matching feature representations of target and reference images. As opposed to previous works, Gatys et al. [24] showed that texture matching in feature space is more effective than pixel-level matching. Although they relate to stain normalization, style transfer works are not directly applicable to histopathology as they rely on matching textures between images and would result in modified tissue structures. These style matching techniques also have the disadvantage of relying on a single reference image to perform the matching and do not account for dataset distributions.

Learning color transformations has also been explored through different techniques for computer vision applications. For instance, Minervini et al. [26], [27] proposed a supervised approach for learning color transformations while maximizing the separability of classes (i.e. objects in an image) in the context of image compression. Most applications of generative learning to color transformations focus on image colorization. For instance, Zhang et al. [28] trained a deep network to map a grayscale input image to an output distribution over quantized color values. These works relate to stain normalization as they learn a transformation of a color space, however, they generally rely on segmentation masks to train the classifier or do not guarantee preserved textures which limits their direct applicability to histopathology.

Our work builds on the foundations of generative adversarial networks (GAN). Adversarial training was introduced by Goodfellow et al. for generative learning [17]. It consists of a generative and a discriminative model trained through an objective function that implements a two-player zero sum game between a discriminator - a function aiming to tell apart real from fake input data, and a generator – a function optimized to generate input data from noise that "fools" the discriminator. The "game" that the generator and the discriminator "play" can be intuitively described as follows. In each step, the generator produces an example from random noise that has the potential to fool the discriminator. The discriminator is then presented with a few real data examples, together with the examples produced by the generator, and its task is to classify them as "real" or "fake". Afterwards, the discriminator is rewarded for correct classifications and the generator for generating examples that fooled the discriminator. Both models are then updated and the next cycle of the game begins.

Inspired by generative learning and style transfer techniques, we model stain normalization as style transfer where image structures must be preserved and matching is to be performed based on statistics over an entire domain of images (not a single reference image). In contrast to previous works in stain normalization, we do not rely on pixel-level matching but use the feature representation of images instead. Note that in our proposed model, normalization is done on the training set which facilitates the application of the trained image analysis model to new test domains. Also, we integrate the mapping as part of the image analysis system at hand, which results in learning class-specific image normalizations which preserve structures as well as color-based predictions.

III. Method

Our goal is to build a discriminative model for a given histopathology image analysis task (e.g., classification or segmentation) with an intrinsic stain normalization component. Such model should be able to handle images with different statistical properties (i.e., different staining appearance) without requiring the need for additional training or pre-processing.

A. Problem Setting

To illustrate our problem, we assume we are given a dataset of images $\{\mathbf{x}^A\}$ from pathology lab A with their corresponding annotations $\{y^A\}$ (e.g., segmentation masks or class labels). We are also given a set of images $\{\mathbf{x}^B\}$ from a second pathology lab B without annotations. We assume images $\{\mathbf{x}^A\}$ and $\{\mathbf{x}^B\}$ have different staining appearance

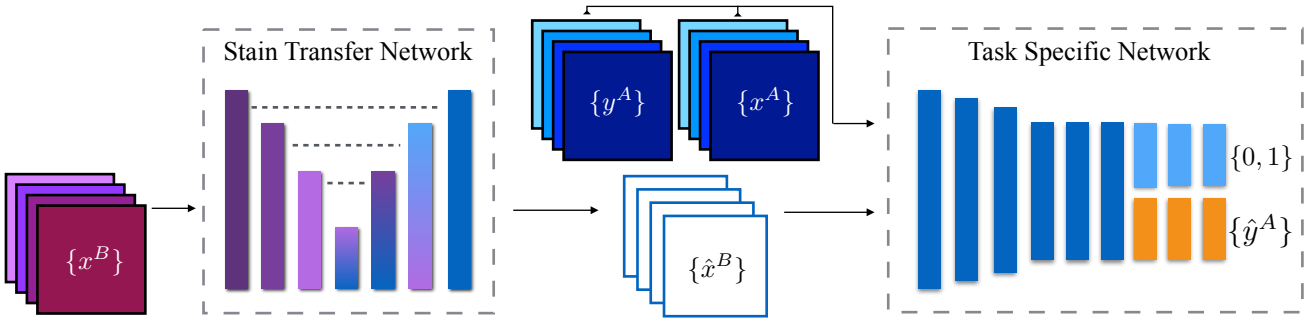

Fig. 2: Proposed architecture. $\{\mathbf{x}^A\}$ are training images and $\{y^A\}$ are their corresponding annotations (e.g., class labels or segmentation masks). $\{\mathbf{x}^B\}$ is a set of test images representing the same tissue types but drawn from a different distribution (e.g., acquired in a different pathology lab with different staining agents or scanned with a different microscope). The stain transfer network is a generative model with an encoder-decoder architecture and skip connections (in dashed gray lines between blocks). This generative network learns to generate images $\{\hat{\mathbf{x}}^B\}$ similar to training images $\{\mathbf{x}^A\}$ in terms of staining appearance and to test images $\{\mathbf{x}^B\}$ in terms of content. A task-specific network simultaneously estimates the probability that an input image came from the training set, $P_1(\mathbf{x})$ rather than the generative network, $P_0(\mathbf{x})$, and predicts a task-specific label $\{\hat{y}^A\}$ given real input images $\{\mathbf{x}^A\}$ and generated images $\{\hat{\mathbf{x}}^B\}$. In this configuration, the task-specific network is trained on images drawn from the training set distribution as well as generated images similar (in terms of stain) to the test set.

due to differences in the acquisition procedure (e.g., pathology center, concentration of staining agent, or type of microscope used), thus are drawn from different data distributions. In this scenario, we aim at building a task-specific discriminative model $C(\mathbf{x};\theta_c)$ parametrized by θ_C trained on annotated images $\{\mathbf{x}^A\}$ that can generalize well to unseen images $\{\mathbf{x}^B\}$. Note that we use the notation $C(\mathbf{x};\theta_c)$ when referring to any input image x^A or x^B. The following presents the details of our model and its learning formulation in the context of image classification, however the model could be easily generalized to other histopathology image analysis applications (e.g. nuclei detection, cell segmentation, ROI delineation).

B. Model Definition

Figure 2 shows the different components of our proposed model. At a high level, given labelled training images $\{\mathbf{x}^A\}$ and unlabelled test images $\{\mathbf{x}^B\}$, we build a task-specific discriminative network equipped with a stain normalization component (i.e., stain transfer network) that performs a non-linear mapping of images from different distributions. Stain normalization is performed in an adversarial setting in which a generative model $G(\mathbf{x}^B;\theta_G)$ (i.e., the stain transfer network) competes against a discriminative model $C(\mathbf{x}^A;\theta_C)$ (i.e., the task specific network) and generates new images $\hat{\mathbf{x}}^B$ that should "fool" C whose task is to discriminate between real images $\{\mathbf{x}^A\}$ and generated normalized training images $\hat{\mathbf{x}}^B$. While identifying real from generated images, the task-specific discriminative model C is simultaneously learning to classify images given real training images $\mathbf{x}^A$ and unlabelled generated images $\hat{\mathbf{x}}^B$, as if training and test datasets were sampled from a common distribution. Models G and C can be any differentiable functions. In our implementation, we chose to use convolutional neural networks given their success in a variety of histopathology image analysis applications.

As opposed to standard stain normalization techniques, we do not rely on a single reference image to learn a color mapping, as this strategy cannot represent the distribution of the entire test set. In fact, the task of the stain transfer network G is to learn a dataset-specific probability distribution such that generated (normalized) images show similar statistics to both training and test sets. Also, in the proposed model, stain normalization is combined with image classification and the full model is trained end-to-end, which results in a class-specific and image-specific stain normalization.

Network Architecture: We design our stain transfer and task-specific networks architectures by adapting state-of-the-art convolutional networks architectures previously applied to histopathology. Both networks use sequential modules of the form: convolution – batch normalization – ReLU.

We used an encoder-decoder architecture for the stain transfer network, in which we included skip connections from encoder to decoder layers as proposed in fully convolutional networks [29]. Skip connections are established between desired pairs of layers and consist of concatenating feature maps (or channels) of a given layer i with the features maps at layer j. The encoder-decoder architecture maps a high resolution input image to a normalized high resolution output image by first down-sampling the image through series of non-linear modules (encoder) then upsampling the coarse encoded input. The addition of skip connections allows us to share low level information between input and output, such as textures and the location of prominent edges. The skip connections are added between each layer i of the encoder and its reverse layer j in the decoder (Figure 2).

The goal of the task specific network is to: (i) discriminate generated image from real images and; (ii) classify images (in case of image classification task). To do so, the task specific network architecture can be defined as any desired network

architecture on which we perform simple modifications of the penultimate layers. We chose to use AlexNet [20]. The three last fully connected layers of the model (in orange in Figure 2) serve as task-specific layers (e.g., classifier). In parallel to the classifier, we add new fully convolutional layers on top of the last pooling layer of the original network, which will serve as stain-specific layers in the adversarial training game. The goal of these additional fully convolutional layers is to identify normalized images $\hat{\mathbf{x}}^B$ from original training images $\mathbf{x}^A$, similar to the discriminator networks in generative adversarial networks [17], using the feature representations of the images. The fully convolutional nature of these layers allows us to discriminate real from generated images using local patch-level information instead of the global image which results in finer stain normalization. In the proposed architecture, early layers of the task-specific network are trained to simultaneously analyze (e.g., classify) images and distinguish generated from real images, thus enforcing robustness towards staining variations from specific distributions in the trained model without totally discarding colors. Intuitively, the architecture of the task-specific network is designed such that the model is forced to learn high-level semantics and low-level stain-specific (or style) information.

Objective: The objective of the model is to generate stained images while learning a specific analysis task. Formally, this can be modeled with the following optimization problem:

$$\min_{\theta_G,\theta_C} \max_{\theta_C} \alpha\mathcal{L}_{\text{adv}}(C,G) + \beta\mathcal{L}_r(G) + \gamma\mathcal{L}_c(G,C) \qquad (1)$$

where α, β and γ are hyper-parameters setting the importance of each term in the optimization problem. θ_C and θ_G are the parameters of the task-specific and stain transfer networks, respectively.

Learning to transfer stains involves playing a minimax game between both networks. The stain transfer network learns to generate images by fooling the classifier whose task is to identify images with different stains. This is captured in the optimization problem by the adversarial loss $\mathcal{L}_{\text{adv}}$. The regularization loss $\mathcal{L}_r$ is used to preserve structures when generating images and the classification loss $\mathcal{L}_c$ is task-specific and allows the model to simultaneously analyze images and learn class-specific color transformations. Below, we elaborate on each of the terms of the objective function.

Transferring Stains: The first part of our stain normalization strategy consists of learning to generate images $\hat{\mathbf{x}}^B$ that preserve the content of the original image $\mathbf{x}^B$ but possess the staining appearance of images $\{\mathbf{x}^A\}$. We refer to this transformation as stain transfer. In a typical stain normalization technique, this step is performed by matching the statistics of $\mathbf{x}^B$ and a reference image $\mathbf{x}^A$ at pixel level using RGB information. Rather than matching RGB values, we learn to generate images with the staining properties of the entire domain of images $\{\mathbf{x}^A\}$. This implies learning the probability distribution of images $\{\mathbf{x}^A\}$, which can be achieved using an adversarial loss function. Specifically, the adversarial loss involves the stain transfer network $G(\mathbf{x}^B;\theta_G)$ that maps an input image $\mathbf{x}^B$ to a stain normalized image $\hat{\mathbf{x}}^B$ and the task specific model $C(x;\theta_C)$ that outputs the likelihood of a given image $\mathbf{x} \in \{\mathbf{x}^A, \hat{\mathbf{x}}^B\}$ to be sampled from the real training set distribution. Formally, the adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_{adv}(C,G) = \mathbb{E}_{\mathbf{x}\sim\{\mathbf{x}^A\}}[\log C(\mathbf{x};\theta_C)] \\ + \mathbb{E}_{\mathbf{x}\sim\{\mathbf{x}^B\}}[\log(1 - C(G(\mathbf{x};\theta_G);\theta_C))]\end{aligned} \qquad (2)$$

where G minimizes the above loss while C maximizes it. At equilibrium, the color statistics of generated images $\{\hat{\mathbf{x}}^B\}$ should match those of real training images $\{\mathbf{x}^A\}$ [17].

Enforcing Structure-Preserved Transformations: To enforce the stain transfer network to preserve tissue structures when learning to generate images, we use an edge-weighted L_2 regularization term that encourages G to preserve salient image edges of the ground truth input. We define the regularization loss as follows:

$$\mathcal{L}_r(G) = \mathbb{E}_{\mathbf{x}\sim\{\mathbf{x}^B\}}||W \circ \mathbf{x} - W \circ G(\mathbf{x};\theta_G)||_2, \qquad (3)$$

where W is the color gradient vector field of the input image and captures the edges of the original input image $\mathbf{x}^B$, and $\circ$ is the Hadamard product (i.e. element-wise multiplication) operator. The multiplication by W focuses the penalty in the difference between the generated and the input images on the edge locations. Ultimately, optimizing G results in preserving the structure of the input image while transforming the staining appearance based on the distribution of images from stain A.

Moreover, we enforce preserved structures with the choice of network architecture for G. In fact, using an encoder-decoder architecture with skip layers enables the information to flow from encoder to decoder at different levels, which results in preserving some of the low level information related to textures (e.g edges).

Learning Class-Conditional Transformations: The stain normalization component is augmented with a task specific model, e.g., classifier in the case of image classification. Given real and transformed training images $\mathbf{x} \in \{\mathbf{x}^A, \hat{\mathbf{x}}^B\}$, the task specific model generates a probability value for each possible labeling. For example, for the classification problem, the task specific model generates probability values for each class. In this case, the classifier's loss function is defined using a cross entropy criterion as follows:

$$\begin{aligned}\mathcal{L}_c(G,C) = \mathbb{E}_{\mathbf{x},y\sim\{\hat{\mathbf{x}}^B,\hat{y}^B\}}[\log C_y(G(\mathbf{x};\theta_G);\theta_C)] \\ + \mathbb{E}_{\mathbf{x},y\sim\{\mathbf{x}^A,y^A\}}[\log C_y(\mathbf{x};\theta_C)]\end{aligned} \qquad (4)$$

where the first expectation corresponds to the classification loss of generated images using pseudo labels $\hat{y}^B$ that correspond to predicted classes for the generated images and the second expectation represents the classification loss for training images. Pseudo labels [30] are recalculated at each training iteration using the model's current parameters. Using pseudo labels during training is a semi-supervised learning strategy that allows us to handle non-annotated images while training the network. It has been shown that pseudo labels generally result in improved generalization performance by favoring a low density separation between classes [30]. By

Algorithm 1: Training algorithm with minibatch stochastic gradient descent

Input : $\{\mathbf{x}^A\}$, $\{y^A\}$ and $\{\mathbf{x}^B\}$
Initialize network's weights.

for *number of training iterations* **do**
Sample minibatch of M images: $\{\mathbf{x}_m^A\}$, $\{\mathbf{x}_m^B\}$
Generate $\{\hat{\mathbf{x}}_m^B\}$ by $G(\mathbf{x}_m^B;\theta_G)$
Update C by ascending in the direction:

$$\nabla_{\theta_C}\Big(\sum_{m=1}^{M}\log C(\mathbf{x}_m^A;\theta_C)+\log\Big(1-C(\hat{\mathbf{x}}_m^B;\theta_C)\Big)\Big)$$

Update C by descending in the direction:

$$\nabla_{\theta_C}\Big(\sum_{m=1}^{M}\log C_{\hat{y}_m^B}(\hat{\mathbf{x}}_m^B;\theta_C)+\log C_{y_m^A}(\mathbf{x}_m^A;\theta_C)\Big)$$

Update G by descending in the direction:

$$\nabla_{\theta_G}\Big(\sum_{m=1}^{M}\log\Big(1-C\Big(\hat{\mathbf{x}}_m^B;\theta_C)\Big)+||W\circ\mathbf{x}_m^B-W\circ\hat{\mathbf{x}}_m^B||_2\Big)$$

end

combining classification with stain normalization, the generative model G learns class-specific mappings, which is a desirable property as, in general, staining should reflect differences between tissue types. Also, this amalgamation of the stain normalization with the classification network retains the discriminative patterns in the normalization process. Note that $\mathcal{L}_c$ can be easily generalized to other analysis tasks. For example, for image segmentation the labels y and predictions $\hat{y}$ would be segmentation masks. The task-specific networks' output would have to be re-sized accordingly.

Algorithm 1 summarizes the training procedure for the proposed method. At each training epoch, we first compute the gradient of $\mathcal{L}_{adv}$ with respect to θ_C and update the parameters by taking a small step in the direction of the gradient (ascending gradient). Then, we compute the gradient of $\mathcal{L}_C$ with respect to θ_C and update the parameters by taking a step in the opposite direction of the gradient (as here the goal is to minimize $\mathcal{L}_C$). Finally, we update θ_G after computing the gradient of $\mathcal{L}_{adv}+\mathcal{L}_r$ and taking in step in the opposite direction of the gradient (gradient descent). We learned the parameters θ_C and θ_G in a single optimization loop involving successive gradient updates for each network using iterations of stochastic gradient descent. The complete model is trained end-to-end using backpropagation.

IV. Validation Experiments

We performed two types of experiments. First, we qualitatively and quantitatively evaluated the quality of stain-normalized images generated using our model; then, we explored the generality of the proposed method by evaluating its performance on different histopathology image analysis tasks and datasets. The results reported in this section were obtained by applying the model on three histopathology datasets. Each dataset was split into a train, validation and test set. The annotated train and validation sets were used to learn the model's parameters and set the hyper-parameters (i.e. learning rate and momentum) and were composed of images sharing similar properties (e.g. digitized with the same microscope). Images from the test set were drawn from a different distribution and their annotations were only used for final performance evaluation. We used the following datasets:

- The public Mitosis-Atypia scoring dataset of breast histology images released as part of the MITOS-ATYPIA ICPR'14 challenge [2]. The dataset consists of 11 histology slides with multiple 20x frames per case scanned with an Aperio scanner and re-scanned with an Hammatsu scanner. To evaluate the performance of the model given inter-microscope variability we used Hammatsu images for training and Aperio images for test. We extracted 250×250 non-overlapping patches per frame and obtained a total dataset of 3360 Hammatsu images for training and 1440 for validation and 4800 Aperio images for test. Note that given the particularity of this dataset, the test set and training set have larger overlapping distributions (in terms of textures and image content but not in terms of staining) than the train and validation sets. Given the number of instances per set, we can expect higher test than validation accuracy.
- We also used the MICCAI'16 GlaS challenge dataset [3], which consists of colon adenocarcinoma tissue images. We used this dataset to test the proposed model on two tasks: benign vs malignant tissue classification and colon gland segmentation. We used 85 images for train, 20 for validation and 60 for test. We extracted 250×250 non-overlapping patches from all images and used texture-only augmentation via elastic warpings [31] to augment the training set.
- Finally, we tested the model on whole slide images of ovarian carcinomas acquired in different histopathology centers and scanned with different microscope types[1]. Each slide represents a unique patient's tumour biopsy to be classified into one of five ovarian carcinoma subtypes (Figure 3). Sixty (60) slides were used for training, 20 for validation and 55 for test. All training and validation images came from the same pathology center while test images were gathered from multiple centers. Each whole slide image was represented by 250×250 non-overlapping patches from the 20x microscope magnification covering the entire tissue slide. Given the large size of these whole slide images (on average $50,000\times50,000$ pixels), we omitted patches that contained more than 60% of background pixels. Examples of the original whole slide images are shown in Figure 3. This dataset has different levels of difficulty. First train and validation images were 'cherry-picked' by a set of expert pathologist, in the context of a reproducibility study, for being most representative of each subtype under 'perfect' acquisition procedures (similar tissue cut, similar stain concentration and similar microscope) and thus strongly differ from test images. Also, the entire dataset reflects the prevalence of the diagnosed subtypes with a significant imbalance towards high-grade serous carcinomas (HGSC).

[1]We made this dataset available at: https://tinyurl.com/jpuxj6g

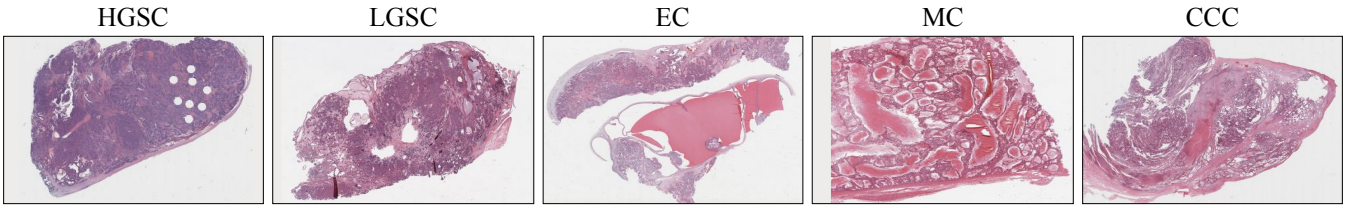

Fig. 3: Ovarian carcinomas subtypes. Acronyms represent the subtype of each whole slide image: High grade serous carcinoma; low grade serous carcinoma; endometriod carcinoma; mucinous carcinoma; and clear cell carcinoma.

Models	MITOSIS			COLON-CLASSIF			COLON-SEGM			OVARY		
	Train	Val.	Test	Train	Val.	Test	Train	Val.	Test	Train	Val.	Test
Without Normalization	96.4±0.1	74.9±0.1	71.5±0.3	99.6±0.2	**90.0±0.1**	81.6±0.1	95.9±0.4	93.5±0.3	68.5±2.2	88.1±2.1	55.1±2.0	44.8±1.5
Grayscale	87.1±0.1	**83.1±0.1**	83.0±0.1	99.5±0.1	85.0±0.1	63.4±0.1	90.5±0.1	86.4±0.1	70.1±0.1	93.1±0.5	51.4±0.6	32.5±1.0
Augmentation	**97.4±0.1**	80.5±3.0	85.9±3.4	**99.9±0.1**	**90.0±1.1**	**87.5±2.1**	95.5±3.0	94.0±1.0	78.5±0.5	**95.5±2.0**	61.0±0.4	35.0±3.8
Color-Matching [9]	96.6±1.0	73.9±1.0	70.1±0.1	99.5±0.1	89.5±0.5	65.5±0.3	95.1±0.4	93.5±1.5	80.1±0.1	86.5±1.0	59.0±0.3	33.7±0.1
$\mathcal{L}_{adv}$	96.1±0.2	74.9±0.1	78.8±0.2	99.5±0.1	**90.0±0.0**	84.6±0.1	95.6±0.1	95.1±0.1	80.2±0.3	88.5±1.0	58.1±1.0	56.5±1.0
$\mathcal{L}_{adv} + \mathcal{L}_r$	96.1±0.2	75.0±0.3	85.9±0.1	99.4±0.1	**90.0±0.0**	85.7±1.1	95.4±0.1	96.2±1.2	80.6±0.4	88.0±0.5	58.0±0.9	59.9±0.4
$\mathcal{L}_{adv} + \mathcal{L}_r + \mathcal{L}_c$	95.1±0.1	76.2±0.3	**90.0**±0.1	88.5±0.1	88.1±0.1	86.7±0.2	**96.2±0.1**	**96.5±0.5**	**82.0±0.9**	92.2±0.8	**68.1±2.3**	**61.7±1.3**

TABLE I: Performance of the proposed stain transfer strategy vs other common approaches. Reported values are in percentages and reflect the average accuracy and variance on different train, validation and test images. The four first rows correspond to the performance of the task-specific network (i.e. AlexNet) on different dataset configurations (images are un-normalized, grayscale, color-augmented or normalized with [9]) without using the proposed stain transfer network. The four last rows report the performance of the proposed model combining the stain transfer network with the task-specific model and the benefit of the different penalty terms used in the proposed multi-loss formulation.

Finally, color inconsistencies not only appear across whole slide images but also within tissue slides. In fact, there are under- and over-stained regions that appear within single whole slide images in the test set, while this is seldom observed in the training set images.

A. Experimental Design

We used a U-Net architecture [31] for the stain-transfer network and AlexNet for the task-specific network. No pre-processing was applied on images. To avoid over-fitting, each convolution layer of the U-Net architecture was followed by batch normalization [32] and we used dropout in the last fully connected layers of AlexNet. We used stochastic gradient descent with a learning rate of 1e-3 to train the task-specific network and 2e-4 to train the stain transfer network. Training hyper-parameters was performed on the validation sets for each dataset. We implemented the model with Torch library [33] and trained it on an NVIDIA Titan X GPU.

We compared our approach to different baselines. First, we tested the influence of color on the performance of the classifier using three experimental settings: without normalizing images with respect to stain variations; using grayscale only images; and using color-based data augmentation [20]. Then, we tested pre-processing all test set images via a popular color matching strategy [9] that relies on learned color deconvolution stain matrices and non-linear color mappings. For this baseline, we randomly selected ten reference images from the training set and applied the trained AlexNet model on the 10 normalized test sets. Finally, as a proof-of-concept of the utility of adversarial training in stain normalization, we tested different combinations of our proposed loss functions: (i) we tested using an adversarial loss only ($\mathcal{L}_{\text{adv}}$) to generate normalized images that are then fed to an independently-trained classifier; (ii) we tested the influence of the regularization loss $\mathcal{L}_r$ combined with the adversarial loss $\mathcal{L}_{\text{adv}}$; and (iii) we tested the full loss function that combines the task-specific loss $\mathcal{L}_c$ with the adversarial loss in an end-to-end framework. Note that all experiments were reproduced twice after swapping the train and test sets such that there was no overlap in the training and validation sets between images from different distributions.

B. Results

We first discuss the quantitative results. Table I shows the averaged performance of the proposed method as well as different baselines on different datasets and tasks in terms of image-level classification accuracy and pixel-level segmentation accuracy. Note that we will use the terminology "original" test dataset when referring to the un-normalized test set and "normalized" test set when referring to stain normalized test images.

Importance of Color: Results shown in Table I confirm that: (i) staining inconsistencies between train and test datasets affect the performance of state-of-the-art image analysis systems (see Table I, "Without Normalization"); (ii) access to color information improves the accuracy (see Table I, "Grayscale"). In fact, discarding color information using grayscale only images or color-based data augmentation improves the generalization ability of the analysis system but only on certain datasets. For

instance, on the MITOSIS dataset, which consists of images with a completely similar content (or texture) but different staining appearance due to the different acquisition microscope, discarding color does boost the generalization of the classifier (test accuracy of 71.5% without normalizing images vs 83.0% when using grayscale images). In contrast, using grayscale images deteriorates the performance of the classifier on test images for both COLON and OVARY datasets (e.g. with grayscale images, test accuracy drops from 81.6 to 63.4% on COLON-CLASSIF and with color-only data augmentation, accuracy drops from 44.8 to 35.0% on OVARY). Color appearance generally plays an important role in identifying subtypes of ovarian carcinomas. For instance, color is critical in identifying HGSC (Figure 3) as it reflects the abundance of nuclei in the tissue image, which is one of the important characteristics of this subtype. This characteristic of ovarian carcinomas may explain the poor performance observed with color-based data augmentation on this dataset. In fact, data augmentation is based on the assumption that objects within images can be categorized regardless of their color appearance, which results in learning color-invariant features. This assumption does not hold in the case of ovarian carcinomas for which color is an important feature. Also, variations in staining often appear within a given ovarian tissue slide thus adding to the inter-slide variations. Stain normalization techniques that rely on color-matching are not designed to handle intra-slide stain variations that exist on whole slide images [34] as they solely focus on the standardization of patch images containing a small region within the whole slide image.

We tested the statistical significance of the results by comparing the test accuracy obtained with our proposed method (last row of Table I) with respect to all other baselines (first four rows of Table I). We used a Wilcoxon signed-rank test and obtained statistically significant differences at $p < 0.05$ on all experiments except on the COLON-CLASSIF dataset when comparing our method to data augmentation. In fact, we observed that using data augmentation on colon classification performs better than other approaches, including ours. The colon adenocarcinoma dataset is the smallest dataset used in our study (only 85 training images). It is unclear whether the performance gain observed using color augmentation is due to the dataset size or to the irrelevance of color information when classifying colon adenocarcinomas. We note nonetheless that the train/val/test accuracy of the proposed model show relatively good generalization on the test set. In contrast, when using data augmentation the model almost perfectly fit (i.e. 100% accurate) to the train set but achieved only 87.5% accuracy on the test set which may indicate bias towards the training set.

Importance of Normalization: Without color normalization and regardless of the task or the dataset, we consistently observed a drop in performance between train, validation and test sets (e.g., -10.3% to -43.3% between train and test sets of whole slide images). This confirms the need for normalization in automatic histopathology image analysis. We can also observe in Table I that the classification performance *on the training and validation sets* when using the proposed

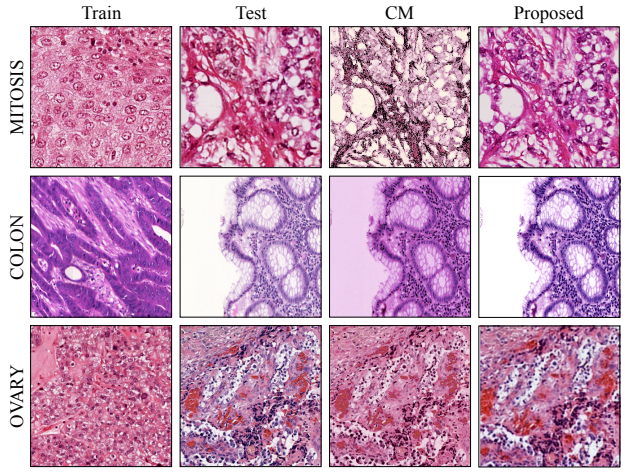

Fig. 4: Examples of stain normalized images. "Train" refers to the image randomly selected as reference for the baseline color-matching ("CM") method [9]. "Test" corresponds to original test images before stain normalization. Normalized images obtained with our proposed method are shown in the last column and were obtained using the entire training set.

method (three last rows of Table I) is comparable to the baseline "Without Normalization" which uses the classifier loss only L_c and thus reflects the performance of the AlexNet model alone (without the U-Net). In contrast, we observed improvements on the *test accuracy*. This confirms that the performance gain is not only the result of the architecture design (combination of U-Net and AlexNet) as this would also reflect in improvements of the train and test performance, but is linked to the intrinsic stain normalization resulting from combining these architectures with the proposed loss function. Note that testing the influence of the combination of both architectures (U-Net and AlexNet) implies using a specific loss function to train the generative model (i.e. U-Net) to generate images; this loss function can be either L_{adv} or $L_{adv} + L_r$.

Color Matching vs Style Matching: We used the color-deconvolution based approach of Khan et al. [9] to normalize test images to a randomly chosen reference image from each training set. We repeated the random selection of a reference image ten times and report the best accuracy results obtained for each dataset. We observed large differences in performance when randomly choosing different reference images for normalization. Test accuracy differed by as much as 38% over the 10 experiments with different randomly chosen reference images. Some reference images resulted in a decrease in accuracy from the test accuracy obtained with un-normalized images (e.g. up to -6% on MITOSIS, -22.8% on COLON and -14.2% on OVARY dataset). These results confirm the sensitivity of color matching techniques to the choice of reference image.

Using the adversarial loss $\mathcal{L}_{adv}$, we trained the stain transfer network G to generate normalized images then applied the trained AlexNet model on normalized test set images gener-

ated with G. Our results (Table I) show that style matching with adversarial training outperforms color matching on all datasets and for all tasks (we observed an increase of 2 to 28% from "without normalization" test accuracy when using L_{adv} instead of color-matching).

Examples of generated images using our proposed method on failure cases of the color matching approach [9] are shown in Figure 4. In general, test images generated with the proposed method are sharper and better match the color distribution of training images. We observed that color-matching techniques are sensitive to the distribution of tissue components in the reference and source image, which may result in erroneous mappings (e.g., COLON case in Figure 4 shows background pixels erroneously colored). Also, as they rely on a single image to learn color transformations, color matching methods are susceptible to the presence of artifacts in the reference image, which can lead to further distancing the normalized images from the training set color distributions (e.g., the MITOSIS and OVARY cases in Figure 4 show over-stained and under-stained normalized nuclei).

Class-Specific Stain Normalization: We investigate the advantage of combining stain normalization and image analysis (full optimization problem in eq.(1)). In Table I, we show the performance of the full optimization problem ($\mathcal{L}_{adv}+\mathcal{L}_r+\mathcal{L}_c$) compared to various baselines. In general, we observed a clear advantage from embedding the normalization with the analysis task at hand as it shows to improve the generalization of the trained model on test images (e.g. +18.5%, +5.1% and +16.9% in classification accuracy on MITOSIS, COLON and OVARIAN datasets when using the full optimization problem vs. no normalization). These results confirm that C learns discriminative features that are more robust to stain variations when using the proposed two-branch architecture trained with the proposed full objective function.

Our model can be seen as a form of guided data augmentation as the task-specific network sees labelled training images as well as unlabelled normalized test images during training. In contrast to other augmentation techniques, in our model, the additional synthetic data provided during training is drawn from a more realistic distribution that corresponds to the learned mixture of training and test set distributions. Our strategy generally performed better than color-based data augmentation using fancy PCA [20] (e.g., +3.5 to 26.7% increase on test accuracy for MITOSIS and OVARY datasets, see Table I). In color-based data augmentation, the generated images are training images with different color appearance while, in our method, the analysis model is provided with synthetic images from the test set.

We also investigated the effect of amalgamating the stain normalization with the image analysis model on the quality of normalized images. To assess the distribution similarity between train and test images as well as train and normalized test images, we computed the average histogram intersection score [35] between train and original test images across classes for all three datasets and compared it with the histogram intersection score between train and test images normalized using the proposed model. The results are reported in Table II. The histogram intersection score measures the similarity between two image histograms computed in Lab color space. Higher scores indicate more similar color distributions.

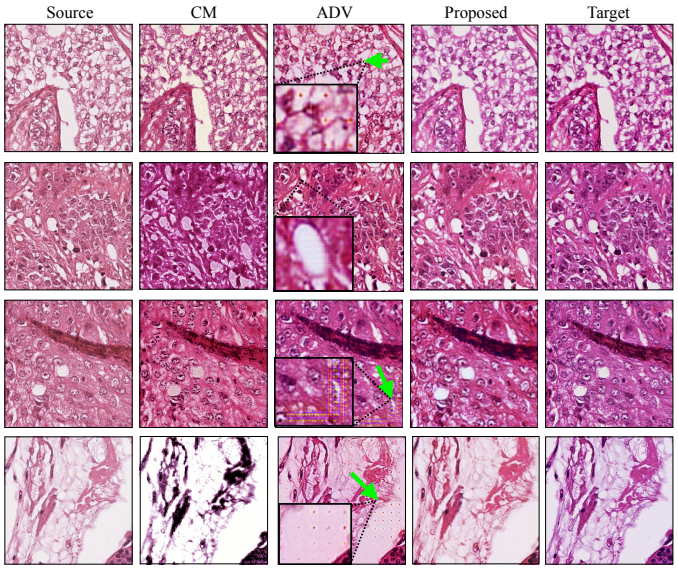

Fig. 5: Examples of stain normalized images on the MITOSIS dataset. "Source" are the original training image, "Target" the corresponding original test image scanned with a different microscope, "CM" is the output of [9] and "ADV" the generated normalized image using G trained with $\mathcal{L}_{adv}$ (without the proposed regularization terms). Red arrows are used to highlight artifacts. "Proposed" correspond to the output of G when trained with the full objective function (eq.(1)).

First, the generally low histogram intersection scores obtained between train and original test images reflect the difficulty of the analysis tasks considered in these datasets. In fact, greater differences in distributions will lead to poor generalization of the analysis system. Note that these scores are averaged over classes and the standard deviation reported in Table II reflects differences between classes.

Normalizing images with our proposed model (i.e. Train-NormTest in Table II) generally results in significantly higher similarity scores which indicates that color distributions of test images have a higher match with those of the training dataset after normalization (e.g. up to 23% increase in histogram matching scores after normalization, see Table II. Overall, our results confirmed that there is a joint benefit in normalizing stains while analyzing images as it first helps training more robust analysis systems but also allows to learn class-specific transformations. The proposed normalization strategy also outperforms color-matching techniques [9] which confirms the benefit of using dataset distributions to normalize stains instead of single arbitrarily chosen target images (see Table II Train-NormTest vs. Train-CM-Test).

We also show examples of the normalized images on the MITOSIS dataset in Figure 5. We compared the quality of the normalized images of a given source image to its corresponding ground truth target image. We trained the

model on the source images and removed their corresponding target images from the test set during training. We compared the color-matching normalization technique [9] that performs a one-to-one mapping from source to target, to our proposed method without and with the regularization term $\mathcal{L}_r$ in the loss function. Overall, we observed a better match between the normalized source images with the proposed method and the real target images. Note that the proposed method does not rely on a one-to-one mapping but learns to transform the domain of source images (training images) to the domain of target images. We also observed that the absence of regularization term can lead to artifacts in the generated normalized images (see green arrows on Figure 5).

Model's Sensitivity: We evaluated the influence of each term of the optimization problem in (1) by sequentially adding them when training the model. Our results showed that using an adversarial loss without regularization can already boost the classification performance on the test set. This is expected because the generative model is conditioned on a given input image, and hence is constrained not to deviate drastically from the input when generating images that can fool C. While adding the regularization loss $\mathcal{L}_r$ during training results in a small classification improvement, it considerably sped up the training time (e.g. 55 training epochs were necessary on average before convergence when using $\mathcal{L}_{adv} + \mathcal{L}_r$ vs 120 when using $\mathcal{L}_{adv}$ only). We were able to increase the performance of all task-specific models on test sets when adding the task-specific loss $\mathcal{L}_c$. In particular, adding $\mathcal{L}_c$ resulted in up to 4% increase in accuracy (Table I) when using a combination of all loss functions proposed in eq.1.

We also tested the effect of hyper-parameters α, and γ in eq.1). We did not observe significant changes in test accuracy when varying α and γ by $\pm 40\%$ and reached a $+5\%$ improvement on the COLON-SEGM task when increasing β while keeping α and γ equally weighted. We observed faster model convergence (only 45 to 50 training epochs until convergence) with β = 2 and α= γ = 1).

Finally, we observed minor test accuracy improvements when using the regularization term on the colon dataset as opposed to the two other datasets. One possible explanation could be the nature of the images in these datasets. Colon adenocarcinomas are composed of recurring glandular objects with pronounced textural properties that differentiate benign from malignant tissues. Malignant glands in adenocarcinomas generally show irregular patterns with dark nuclei forming the epithelial border while benign glands have circular and regular appearance [3]. In contrast, mitosis and ovarian cancer tissues share common textural characteristics between different tissue types (Figure 3). These texture-related characteristics could explain the regularization term being more critical when training models to classify ovarian or mitosis tissues, as preserving textures on normalized images may contribute to the classification task.

Channels	Dataset	Train-Test	Train-NormTest	Train-CM-Test
L	MITOSIS	0.24±0.2	0.32 ±1e-3	**0.51±0.4**
	COLON	0.47±0.2	**0.70±1e-3**	0.09±0.06
	OVARY	0.66±0.3	**0.69±1e-4**	0.03±1e-3
a	MITOSIS	0.47±0.2	**0.70±1e-4**	0.30 ±0.2
	COLON	0.93±0.02	**0.95±1e-4**	0.06 ±0.04
	OVARY	0.24±0.05	**0.26±0.2**	0.23 ±0.1
b	MITOSIS	3e-5±2e-3	**0.22 ±1e-3**	0.21±0.2
	COLON	0.82±2e-3	**0.89±1e-4**	0.30±0.1
	OVARY	0.26±0.3	**0.35±1e-4**	0.05±0.01

TABLE II: Histogram intersection matching between train images and original test images (Train-Test), train images and normalized test images using the proposed approach (Train-NormTest) and train images and normalized test images with the color-matching approach of [9] (Train-CM-Test).

V. Discussion and Conclusions

Stain normalization is paramount to accurate and generalizable histopathology image analysis systems. However, accurately normalizing images with respect to stain is challenging due to the variety of possible sources of staining inconsistencies. In this work we argued that stain normalization techniques should have the following properties: (i) they should involve matching the entire domain distributions rather than matching the statistics of individual images; (ii) they should consider specific tissue characteristics thus should be coupled with image analysis systems; and (iii) they should not alter tissue structures.

We designed a model that combines the above properties by leveraging adversarial examples and deep learning architectures. Using an adversarial loss, our model learns to synthesize images with a specified stain. Stain transfer is learned using the entire set of available images, which results in generated images sampled from a mixed distribution of train and test images. Also, the model jointly analyzes and normalizes images, which results in learning class-specific stain transfers via color-invariance towards class-specific color distributions in the analysis system. Finally, the system is enforced to preserve structures when normalizing images using an image content-based regularized optimization problem. Our model was trained end-to-end with standard gradient descent. In practice, the system could be used in situations where images from a new domain (e.g. acquired in a different pathology center) become available without annotations. A simple fine-tuning of the stain transfer network using the new test images along with the previously available annotated training images would allow to normalize the training images with respect to the new domain distribution, facilitating their analysis using the trained task-specific network.

Validation on three different datasets, including large scale whole slide images, demonstrates the advantage of using the proposed model in image classification and segmentation tasks. Our results showed that normalizing images based on dataset distributions is less sensitive than using standard color-matching strategies (c.f. Table I). Also, we observed that combining classification and normalization helps training more robust classification systems but also learning class-specific stain transformations that do not alter color properties of different tissue classes (c.f. Table II). Finally our results also

confirmed that adversarial training allows us to preserve tissue structures when learning normalizing images as it enforces the model to identify content from stain-specific information.

To the best of our knowledge, this was the first approach involving a fully trainable joint normalization and image analysis model and leveraging adversarial examples. Despite the promising results, the proposed method is not without its limitations. First, while the method does not rely on any pre-processing step, it involves training deep networks and thus relies on the availability of training data to accurately estimate dataset distributions. Also, we did not tailor the architecture of our model to the datasets and tasks used in this study but rather used high-level intuitions based on the desired properties of the model, however there can be many different architectures one can adopt that may result in better performance. For instance, a possible extension to the proposed model could be to share parameters between the two networks (task specific and stain transfer network) which can reduce the number of trainable parameters but also may facilitate learning mixed color distributions from the train and test sets.

Future works will involve further exploring the potential of the proposed method with other network architectures, guiding the learning of the hyper-parameters of the objective function (eq.(1)) by leveraging the model's uncertainty (as proposed in previous work [36]), testing the applicability of adversarial training in digital staining (transferring stains across images stained with different staining agents) and quantifying the ability of expert pathologists in detecting stain-transferred synthesized images.

Acknowledgements: We gratefully acknowledge NVIDIA Corporation for GPU donation and The Natural Sciences and Engineering Research Council of Canada (NSERC) for partially funding this work.

REFERENCES

[1] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," *Annual Review of Pathology: Mechanisms of Disease*, vol. 8, pp. 331–359, 2013.

[2] L. Roux, D. Racoceanu, F. Capron, J. Calvo, E. Attieh, G. Le Naour, and A. Gloaguen, "Mitos & atypia," *Image Pervasive Access Lab (IPAL)*, 2014.

[3] K. Sirinukunwattana *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Medical image analysis*, 2017.

[4] J. D. Bancroft and M. Gamble, *Theory and practice of histological techniques*. Elsevier Health Sciences, 2008.

[5] H. O. Lyon, A. De Leenheer, R. Horobin, W. Lambert, E. Schulte, B. Van Liedekerke, and D. Wittekind, "Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains and chromogenic reagents," *The Histochemical Journal*, vol. 26, no. 7, pp. 533–544, 1994.

[6] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities."

[7] D. Magee, D. Treanor, D. Crellin *et al.*, "Colour normalisation in digital histopathology images," in *MICCAI Workshop On Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy*, vol. 100. Springer, 2009.

[8] M. Macenko, M. Niethammer, J. Marron, D. Borland *et al.*, "A method for normalizing histology slides for quantitative analysis," in *ISBI*, 2009, pp. 1107–1110.

[9] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.

[10] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[11] B. E. Bejnordi, N. Timofeeva, I. Otte-Höller *et al.*, "Quantitative analysis of stain variability in histology slides and an algorithm for standardization," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2014, pp. 904 108–904 108.

[12] A. Basavanhally and A. Madabhushi, "Em-based segmentation-driven color standardization of digitized histopathology," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2013, pp. 86 760G–86 760G.

[13] P. A. Bautista, N. Hashimoto, Y. Yagi *et al.*, "Color standardization in whole slide imaging using a color calibration slide," *Journal of pathology informatics*, vol. 5, no. 1, p. 4, 2014.

[14] A. C. Ruifrok, D. A. Johnston *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, 2001.

[15] A. Janowczyk, A. Basavanhally, and A. Madabhushi, "Stain normalization using sparse autoencoders (stanosa): Application to digital pathology," *Computerized Medical Imaging and Graphics*, 2016.

[16] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, 2016.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu *et al.*, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[18] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi *et al.*, "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.

[19] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1400–1411, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," pp. 1097–1105, 2012.

[21] E. L. Clarke and D. Treanor, "Colour in digital pathology: A review," *Histopathology*, 2016.

[22] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni *et al.*, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[23] D. Onder, S. Zengin, and S. Sarioglu, "A review on color normalization and color deconvolution methods in histopathology," *Applied Immunohistochemistry & Molecular Morphology*, vol. 22, no. 10, pp. 713–719, 2014.

[24] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[25] L. A. Gatys *et al.*, "Image style transfer using convolutional neural networks," pp. 2414–2423, 2016.

[26] M. Minervini, C. Rusu, and S. A. Tsaftaris, "Unsupervised and supervised approaches to color space transformation for image coding," in *ICIP*, 2014, pp. 5576–5580.

[27] ——, "Computationally efficient data and application driven color transforms for the compression and enhancement of images and video," in *CVIE*. Springer, 2015, pp. 371–393.

[28] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*. Springer, 2016, pp. 649–666.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," pp. 3431–3440, 2015.

[30] Y. Grandvalet and Y. Bengio, "Entropy regularization," *Semi-supervised learning*, pp. 151–168, 2006.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[33] R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: a modular machine learning software library," Idiap, Tech. Rep., 2002.

[34] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. van der Laak, "Stain specific standardization of whole-slide histopathological images," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 404–415, 2016.

[35] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *ICIP*, vol. 3. IEEE, 2003, pp. III–513.

[36] A. BenTaieb and G. Hamarneh, "Uncertainty driven multi-loss fully convolutional networks for gland analysis." in *MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis*, vol. 10552. Springer, 2017.