Cross-Institute Histopathology Image Stain Normalization with Deep Convolutional Neural Networks

Chris Lin
Department of Statistics
Stanford University
Stanford, CA
clin17@stanford.edu

Qian (Sarah) Mu
Department of MS&E
Stanford University
Stanford, CA
sarahmu@stanford.edu

Abstract

Histopathology images are microscopic medical images that contain stained whole-slide tissues. It is widely used to observe pathological manifestations such as cancerous cell developments. Diagnosis with automated image recognition is extremely time efficient. However, since there is no standardized staining procedure, each institute yields disparate staining appearances that pose a difficulty for accurate classification. Our study aims to solve the problem of stain normalization by transforming stained images to grayscale and then to stained images that have color distributions closer to the images from a reference institute. We applied Encoder-Decoder and Generative Adversarial Network (GAN) to generate reference stained images from grayscale images of other institutes. With the Encoder-Decoder, the results show that the generated (normalized) images have color distributions closer to the reference images while preserving source tissue structure, suggesting that the model is effective at stain normalization.

1. Introduction

In histopathology, whole-slide images (WSIs) of tissues are studied to detect the presence, localization, and grading of diseases, particularly cancers. Staining is a common practice in histopathology to enhance the contrast and intensity of tissues in bright-field microscopy. In general, the stains hematoxylin and eosin (H&E) are used to effecitvely differentiate the nucleus and cytoplasm. However, variations in specimen preparation, staining temperature, staining solutions, cell fixation, and imaging device settings could cause differences in the WSIs obtained from different medical institutes [1]. Therefore, different medical institutes tend to produce WSIs with varying color distribution and intensity. Although human experts on histopathology can take into account these variations during diagnosis, the performance computer-aided detection (CAD) systems of seems to suffer from these variations. Studies have shown that applying stain normalization can lead to higher performance in CAD systems [2, 3]. For intance, Ciompi*et al*. shows that stain normalization for H&E colorectal cancer images improve the performance on tissue classification of a convolutional neural network (CNN) by up to 20% [3].

In this study, we apply deep CNN to the task of histopathology stain normalization across different institutes. Since WSIs from every medical institute can be transformed to grayscale images, stain normalization can be thought as a mapping from the grayscale image space to the stained color image space of a reference institute. With this hypothesis, we trained an Encoder-Decoder network to transform grayscale histopathology images to stained color images in RGBA. We also trained a Generative Adversarial Network (GAN) to generate stained color images in RGBA with grayscale images as inputs. Compared to the ground truth stained images of the reference institute, the Encoder-Decoder has an average mean square error (MSE) of 17.538, and the GAN has an average MSE of 50.367. Quantitative and qualitative analyses comparing histopathology images from 4 source institutes before and after stain normalization using the Encoder-Decoder show that the normalization indeed shifts the source color distribution closer to the reference color distribution.

2. Related Work

Previous work related to the study of color normalization for histopathological slides fall briefly into three categories: color matching, color deconvolution and generative models.

2.1. Color Matching

The method of color matching establishes a template image, from which the colors are referenced. The input image is matched with the template image as a histogram to produce similar color distribution. Color matching is per-

formed on each RGB channel separately and as a whole. To address the problem that the colors are correlated because of the staining, the stains and tissues are grouped into classes, and color matching is performed on each class separately and the weighted results are taken in the end. Reinhard *et al*. proposed a color matching method that transforms RGB images to CIELAB representation [4]. However, it overlooks the different contribution of tissues in different slices. Supervised color matching involves principle componenent analysis that groups images into clusters using matrix factorization [5]. While unsupervised color matching uses optimization techniques. Training a naive Bayes classifier uses weighted average by having prior knowledge of the stain vector [6].

2.2. Color Deconvolution

Stain deconvolution is a specific method used in microscopy imaging to produce a generic color representation in order to reflect the concentration of the stains by quantifying the RGB values through pigment segmentation. It first takes a staining matrix from deconvolution of the staining channels in, then performs normalization on each channel respectively. Stain deconvolution makes the assumption that the absorbance of the stain and its concentration is linearly related. However, since most of the diagnostic procedures are performed under polychromatic instead of monochromatic conditions, the assumption does not hold. Haub *et al*. [7].

2.3. Generative Models

Our work focuses on generative models. There have been broad applications on color normalization in computer vision, mostly aiming to transfer source images so that they will match the style of a given template, at the same time preserving the characteristics of the target image. In the context of histopathological imaging, most work uses conditional GAN, since it is important to preserve the original structure and pattern of the images. Related work in this area includes stain style transfer [8]. Some researches incorporate the classification of tumorous cells as a discriminator because histopathological imaging needs to identify the counts of different cell types. Bautista *et al*. developed a color correction matrix to convert the pixel colors with color calibration slide [1]. However, supervised learning method are not generic and may not lead to optimal objective value. The tasks of such approaches are normally divided into tissue segmentation and classification, in which a better classification accuracy is the aim after normalization [9]. Most characteristic approaches are training adversarial nets in which the generative model and discriminative model are in competition, resembling a two-player minimax game. [10]

3. Data

3.1. Data Source

The WSIs were obtained from the publicly available CAMELYON17 data set [11]. The data set contains 1,000 WSIs of hematoxylin and eosin (H&E) stained lymph node sections of breast cancer patients, including slides with tumor and normal cells. The data set was provided by 5 medical institutes in the Netherlands. The WSIs from Laboratorium Pathologie Oost-Nederland in Hengelo were considered reference images for stain normalization, and hereby the laboratory is referred to as the reference institute. The other medical institutes are Radboud University Medical Center in Nijmegen (source institute 1), Canisius-Wilhelmina Hospital in Nijmegen (source institute 2), University Medical Center Utrecht (source institute 3), and Rijnstate Hospital in Arnhem (source institute 4). Each institute provided MSIs of 40 patients, with 5 MSIs per patient. The MSIs were converted to TIFF (Tagged Image File Format) in RGBA using the file converter of the open-source ASAP package [12].

3.2. Data Processing

Due to the large file size of WSIs, a common practice is to extract patches from WSIs to accommodate computational constraint. We extracted patches of 256×256 pixels from each WSI and stored them in PNG (Portable Network Graphics) format. Patches with more than 95% white pixels, 95% black pixels, or 90% transparent pixels were removed. From the reference institute, 80,000 patches were randomly selected as the training set, 10,000 as the validation set, and 10,000 as the test set. To study cross-institute stain normalization, 2,500 patches were randomly selected from each source institute.

Hypothesizing that WSIs from different institutes share the same grayscale space, and stain normalization is the mapping from the grayscale space to the stain color space of the reference images, we transformed all patches to grayscale images using the formula $L = 0.299 \times R + 0.587 \times G + 0.144 \times B$ with the Python Imaging Library (version 1.1.7). Examples of RGBA and grayscale patches from all the institutes are shown in Figure 1. All images were scaled to range from -1 to 1 before as input for the neural networks.

4. Methods

4.1. Encoder-Decoder convolutional network

We hypothesize that WSIs from different institutes share the same grayscale space, and stain normalization is the mapping from the grayscale space to the reference stained color space. Therefore, the task of stain normalization can be framed as an image-to-image translation problem (from

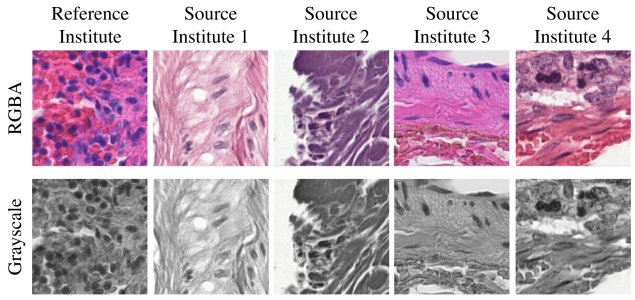

Figure 1. Examples of original RGBA and grayscale patches from the reference and source institutes.

grayscale to stained color images). Zhang *et al*. has shown that an Encoder-Decoder convolutional network achieves good results with grayscale-to-color image translation [13]. In an Encoder-Decoder network, high-dimensional data are encoded into low-dimensional representations through a series of convolution and downsampling layers in the Encoder. The low-dimensional representations are then decoded into high-dimensional representations through deconvolution and upsampling layers in the Decoder [14].

Previous studies have shown that symmetric Encoder-Decoder architectures with skip connections, where encoded outputs are combined with the mirroring decoder inputs during decoding, preserve original image structure and improve image translation quality [15, 16, 17]. Since tissue structure is important for histopathology, we adopted the Encoder-Decoder architecture with skip connections proposed by Isloa *et al*. [15]. The Encoder consists of 8 convolution layers followed by batch normalization and the leaky Rectified Linear Unit (ReLU) [18, 19]. The Decoder consists of 7 deconvolution layers followed by batch normalization and the ReLU activation [20], and a last convolution layer of stride 1 with the tanh activation. The details of the Encoder-Decoder architecture are shown in Figure 2.

4.1.1 Encoder-Decoder objectives

The most common objective of an Encoder-Decoder is the L_2 distance, formalized as

$$L_2(G(x), y) = \mathbb{E}_{x,y}[\|G(x) - y\|_2] \quad (1)$$

where G is the Encoder-Decoder, x is a grayscale image, y is the ground truth stained color image of x, and $\|\cdot\|_2$ is the l_2 Euclidean norm.

It has been suggested that the L_2 objective causes blurring in translated images, and the L_1 distance could alleviate this problem [15]. Therefore, we also consider the L_1 distance as another objective, formalized as

$$L_1(G(x), y) = \mathbb{E}_{x,y}[\|G(x) - y\|_1] \quad (2)$$

where G, x, and y are defined as in Equation 1, and $\|\cdot\|_1$ is the l_1-norm.

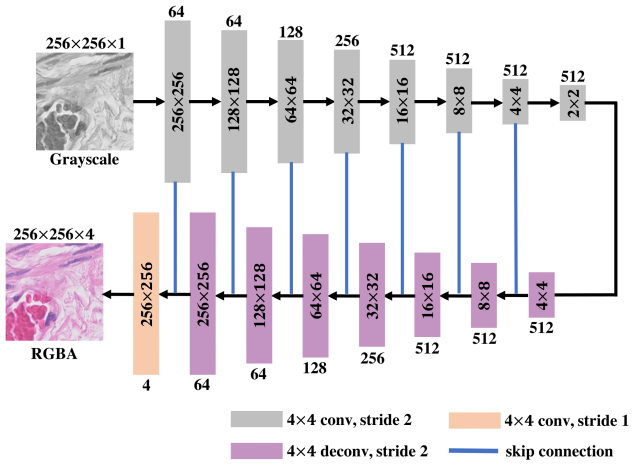

Figure 2. Encoder-Decoder architecture. The Encoder convolution layers (gray) are followed by batch normalization and the leaky ReLU. The Decoder deconvolution layers are followed by batch normalization and the ReLU. The last layer of the Decoder is followed by the tanh activation to output an image of shape $256 \times 256 \times 4$.

4.2. Generative Adversarial Network (GAN)

Stain normalization can also be framed as a generative process, in which a model generates the reference stained color image given a grayscale image. One recent development in generative models is Generative Adversarial Network (GAN) [10]. In a GAN, the Generator network receives samples from a prior distribution of random variable to generate images. The Discriminator network receives generated and real-world images and learns to discriminate the two types. The optimization of the Generator is to maximize the probability of the Discriminator making a mistake, whereas the Discriminator is optimized to minimize this probability. These procedures correspond to a minimax optimization in which the Generator learns to generate images that resemble real-world images. Prior studies have generated color images from grayscale images using GANs [15, 21]. Some recent studies have developed GANs to normalize histopathology stain color [8, 9, 22, 23]. In this study, we also explore applying GAN to stain normalization. In our GAN, the Generator generates reference stained color images, and the Discriminator discriminates generated reference stained color images from real-world reference stained color images.

Instead of using a specified probability distribution as the prior for the Generator, we postulate the grayscale image data distribution as the prior in our GAN, similar to the framework of conditional GAN in [15]. In this way, our Encoder-Decoder was naturally used as the Generator in our GAN. In stain normalization, the Discriminator needs to incorporate information about the grayscale image when discriminating generated and real-world reference stained

color images. Therefore, stained color images are combined with the corresponding grayscale images as inputs to the Discriminator. Our Discriminator is similar to that in [15] and consists of 4 convolution layers: the first convolution layer is followed by the leaky ReLU, and the 3 other convolution layers are followed by batch normalization and the leaky ReLU. The last layer is a fully connected dense layer to output a score. The details of the Discriminator architecture are shown in Figure 3.

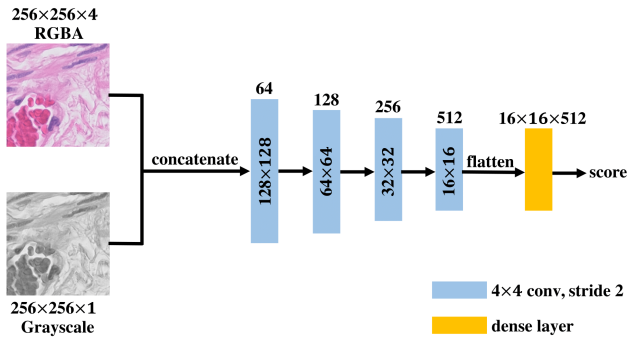

Figure 3. Discriminator architecture. The first convolution layer is followed by the leaky ReLU. The other three convolution layers are followed by batch normalization and the leaky ReLU. The last layer is a fully connected dense layer that outputs a score.

4.2.1 GAN objectives

The Generator objective in the original GAN [10] can be expressed as the following Generator objective:

$$L_G(G(x), D) = -\mathbb{E}_{x\sim p(x)}[logD(G(x))] \tag{3}$$

where G is the Generator (Encoder-Decoder in this study), D is the Discriminator, x is a grayscale image, and $p(x)$ is the distribution of the grayscale image data space. The Discriminator objective can be expressed as:

$$\begin{aligned} L_D(G(x), D, y) = &- \mathbb{E}_{y\sim p(y)}[logD(y)] \\ &- \mathbb{E}_{x\sim p(x)}[log(1 - D(G(x)))] \end{aligned} \tag{4}$$

where G, D, x, and $p(x)$ are as defined in Equation 3, y is the reference stained color image corresponding to x, and $p(y)$ is the data distribution of the reference stained color image.

4.2.2 GAN objectives with image loss

Previous approaches found that when paired data are available the performance of a GAN is improved by incorporating some image loss in the Generator objective [24]. Since we have both the grayscale and stained color images from the reference institute, in this study we explore the effect of adding the L_2 and L_1 image losses to the Generator objective.

We have the following Generator objective for the GAN with additional L_2 loss:

$$\begin{aligned} L_{G,2}(G(x), D, y) = &- \mathbb{E}_{x\sim p(x)}[logD(G(x))] \\ &+ \lambda_2 \cdot \mathbb{E}_{x,y}[\|G(x) - y\|_2] \end{aligned} \tag{5}$$

where λ_2 is a real positive number and a hyperparameter for the importance of the L_2 image loss.

The Generator objective with additional L_1 loss is the following:

$$\begin{aligned} L_{G,1}(G(x), D, y) = &- \mathbb{E}_{x\sim p(x)}[logD(G(x))] \\ &+ \lambda_1 \cdot \mathbb{E}_{x,y}[\|G(x) - y\|_1]. \end{aligned} \tag{6}$$

where λ_1 is a real positive number and a hyperparameter for the importance of the L_1 image loss.

With an additional image loss, the task of the Discriminator remains the same, while the Generator not only has to fool the Discriminator but also generates stained color images close to the ground truths.

4.3. Optimization and inference

All networks were initialized with the Xavier initialization and trained with the Adam optimization algorithm to minimize the objectives [25, 26]. Minibatches of size 16 were used to fit the largest possible minibatch into graphics processing unit (GPU) memory, which reduces as much variance at training time as possible. Cross-validation was used for hyperparameter search to minimize the mean square error (MSE) in RGBA scale (0-255) on the validation set. For all the networks, the optimal learning rate is 0.0002, momentum parameters β_1=0.5, and $\beta_2 = 0.999$. The GAN with L_1 image loss has an optimal λ_1=1.6×10^4. The GAN with L_2 image loss has an optimal λ_2=1.6×10^4. The GANs were trained by first optimizing the Discriminator and then the Generator as proposed in [10].

The Encoder-Decoder with the L_2 objective was trained for 27 epochs, the Encoder-Decoder with the L_1 objective for 26 epochs, the GAN without any image loss for 62 epochs, the GAN with the L_2 image loss for 27 epochs, and the GAN with the L_1 image loss for 28 epochs. All networks were trained for an additional 10 epochs to confirm performance saturation. All networkds were implemented with TensorFlow 1.17.0 and trained on NVIDIA Tesla V100 and NVIDIA Tesla P100 GPUs.

At inference time, the update to batch normalization was frozen and the training data statistics were used. The generated images were converted to RGBA scale (0-255) for evaluation.

5. Experiments

5.1. Image similarity metrics

We use 5 image similarity metrics to evaluate the performance of our models.

Model	MSE	SSIM	PSNR	Pearson Correlation	Histogram Intersection
Encoder-Decoder (L_2)	$\mathbf{17.538 \pm 12.774}$	0.976 ± 0.009	$\mathbf{36.384 \pm 2.539}$	$\mathbf{0.995 \pm 0.004}$	0.925 ± 0.058
Encoder-Decoder (L_1)	18.561 ± 13.722	$\mathbf{0.977 \pm 0.010}$	36.065 ± 2.168	$\mathbf{0.995 \pm 0.004}$	0.902 ± 0.099
GAN	50.367 ± 50.494	0.940 ± 0.016	31.110 ± 2.262	0.942 ± 0.014	0.676 ± 0.059
GAN + L_2	19.201 ± 13.800	0.976 ± 0.010	35.926 ± 2.218	0.994 ± 0.004	0.918 ± 0.068
GAN + L_1	20.072 ± 13.425	0.970 ± 0.011	35.720 ± 2.206	0.994 ± 0.005	$\mathbf{0.940 \pm 0.047}$

Table 1. Model performance on image similarity metrics. Stained color images generated from grayscale images were compared to ground truth stained color images. All images used for the calculations were from the test set of the reference institute.

- The MSE measures the average squared difference between pixels of two images, with a value of 0 for identical images and higher values for more distinct images.
- The Structural Similarity Index (SSIM) measures the structural differences between two images that are most perceivable by the human vision, with a value of 1 for identical images and lower values for more distinct images [27].
- The Peak Signal-to-Noise Ratio (PSNR) is the maximum ratio between the signal and differential noise between two images. The PSNR has a value range of $(-\infty, \infty)$, with higher values for more similar images.
- The Pearson Correlation measures the magnitude and direction of the linear relationship between the pixels of two images. A more positive Pearson Correlation value indicates more similar images [28].
- The histogram intersection is the fraction of overlap between the color histograms of two images, with a value of 1 for identical images and lower values for more distinct images.

5.2. Analysis of models and objective functions

The models were evaluated with the test set from the reference institute. The image similarity metrics were calculated for each pair of generated stained color image and the corresponding ground truth image. The average and standard deviation of the similarity metrics for each model are reported in Table 1. No serious overfitting was observed as the average MSE of the training set is lower than the average MSE of the test set by a value of 1-5 for all the models.

The Encoder-Decoder with the L_2 objective has the best performance in terms of MSE, PSNR, and Pearson Correlation. The Encoder-Decoder with the L_1 objective (average SSIM=0.977) performs marginally better than the Encoder-Decoder with the L_2 objective (average SSIM=0.976). The GAN with L_1 image loss (average histogram intersection=0.940) has the best performance in terms of histogram intersection, slightly better than the Encoder-Decoder with the L_2 objective (average histogram intersection=0.925).

The GAN without any image loss has the worst performance in all similarity metrics. The GANs with L_2 and L_1 image losses have comparable performances to the Encoder-Decoders. However, since the optimal values of $_2$ and λ_1 are large, it suggests that the Encoder-Decoder architecture is more suited for stain normalization than the GAN.

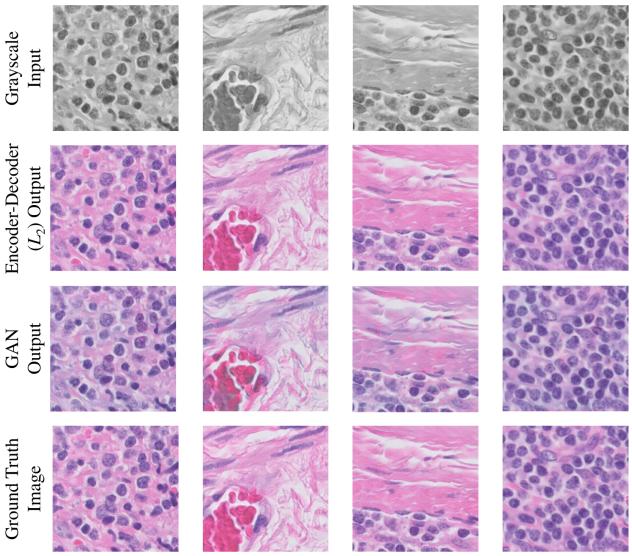

Figure 4. Randomly selected sample outputs of the Encoder-Decoder (with L_2 objective) and GAN. Images are from the test set of the reference institute.

5.3. Error analysis on failure cases

Overall, the Encoder-Decoder with the L_2 objective has the best performance, and the GAN without any image loss has the worst performance. Here we study the outputs of these two models. Both the Encoder-Decoder and GAN seem to be able to generate stained color images that resemble the ground truth images (Figre 4). In particular, the Encoder-Decoder generates stained color images that appear identical to the ground truth images to the human eye, preserving the tissue structure and learning the reference color distribution. This agrees with the fact that the Encoder-Decoder with the L_2 objective has lowest average

Source Institute	MSE	SSIM	PSNR	Pearson Correlation	Histogram Intersection
Institute 1	120.130 ± 67.499	0.957 ± 0.011	28.118 ± 2.767	0.974 ± 0.008	0.806 ± 0.107
Institute 2	165.530 ± 123.73	0.956 ± 0.013	27.569 ± 4.082	0.976 ± 0.010	0.778 ± 0.086
Institute 3	75.105 ± 78.131	0.958 ± 0.018	30.712 ± 3.205	0.977 ± 0.013	0.830 ± 0.064
Institute 4	145.452 ± 96.372	0.955 ± 0.013	27.361 ± 2.692	0.967 ± 0.013	0.774 ± 0.119

Table 2. Encoder-Decoder (L_2) performance on cross-institute stain normalization. Normalized stained color images generated from grayscale images were compared to source stained color images for each institute.

MSE. The stained color images generated by the GAN seem to have areas that are more purple than the ground truth images (Figure 4). This suggests that the Discriminator might be fooled by more purple images, as the WSIs from the reference institute tend to be more purple.

We also examine the generated stained color images that are dissimilar to the ground truth images with respect to each similarity metric. As shown in Figure 5, the Encoder-Decoder seems to fail when the ground truths have colors not commonly seen in WSIs such as black and blue. Also, the Encoder-Decoder seems incapable of reconstructing tissue structures that are out of focus from grayscale images, as suggested by the case with low SSIM (Figure 5). The GAN seems to fail when the ground truth images have areas out of focus or blurring (Figure 6). The GAN also generates colors close to green, which is uncommon in WSIs.

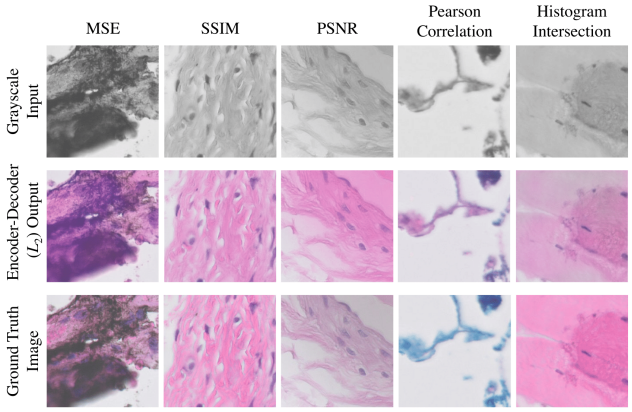

Figure 5. Sample failure cases of the Encoder-Decoder (with L_2 loss). Images from the test set of the reference institute with high MSE, low SSIM, low PSNR, low Pearson Correlation, and low histogram intersection are shown in the corresponding columns.

5.4. Analysis on cross-institute stain normalization

Finally, we study how the Encoder-Decoder with the L_2 objective performs on the task of cross-institute stain normalization. Since we do not have the reference staining for the source institutes we analyze the stain normalization performance by comparing the source images before and after normalization.

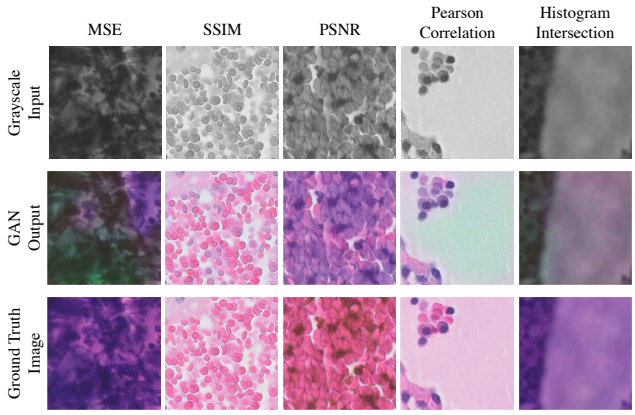

Figure 6. Sample failure cases of the GAN. Images from the test set of the reference institute with high MSE, low SSIM, low PSNR, low Pearson Correlation, and low histogram intersection are shown in the corresponding columns.

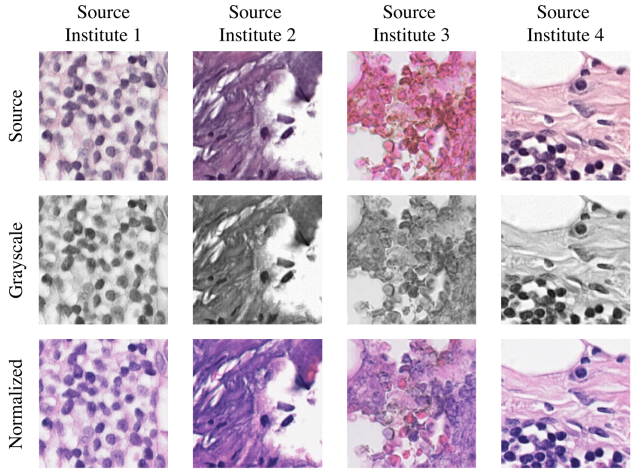

Figure 7. Randomly selected sample images normalized by the Encoder-Decoder (with L_2 objective) from source institutes 1, 2, 3, and 4.

The stain normalization seems to work as the normalized source images appear more similar to the reference images after normalization (Figure 7). When we compare the source images before and after stain normalization, the average SSIM, PSNR, and Pearson Correlation are relatively high, suggesting that the tissue structures were preserved

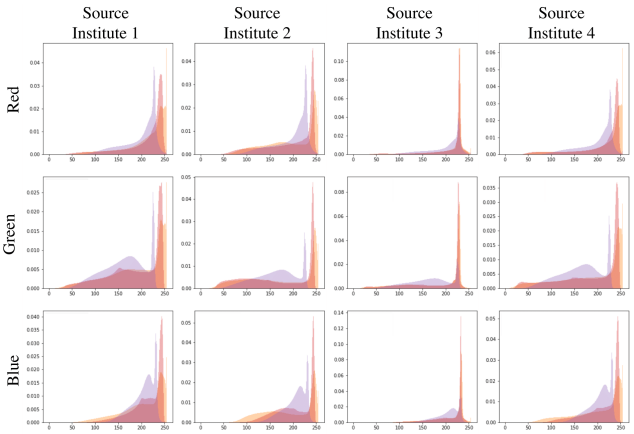

Figure 8. Comparison of RGB color histograms before and after normalization. Purple: color histograms of images from the test set of the reference institute. Orange: color histograms of source images from the source institutes. Red: color histograms of normalized images from the source institutes. Note that the vertial axes have different ranges.

after normalization. The average MSE is large and the average histogram intersection is low, indicating that the color distributions of the source images were shifted after normalization (Table 2). Indeed, the RGB color histograms of the source images were shifted closer to the color histograms of the reference images after normalization (Figure 8). These results suggest that the Encoder-Decoder is effective at the task of cross-institute stain normalization.

6. Conclusion

In this study, we apply Encoder-Decoder and GAN to the task of cross-institute histopathology stain normalization. We hypothesize that stain normalization can be framed as a mapping from source stained images to grayscale images and then to reference stained images. With this hypothesis, we trained Encoder-Decoder networks and GANs to transform any grayscale images to stained images with the reference color distribution. And Encoder-Decoder with the L_2 objective achieved the best performance in terms of MSE, PSNR, and Pearson Correlation when the generated images were compared to ground truth images from the reference institute. The GAN without any image loss performs the worst in all similarity metrics. This suggests that when ground truth stained images are available, directly optimizing the image loss is more effective than training a GAN.

Images from 4 different source institutes were compared before and after normalization by the Encoder-Decoder. Quantitative and qualitative analyses show that the normalized images have color distributions shifted closer to the reference images, while the tissue structures were maintained.

For future works, we intend to study the effect of stain normalization using the Encoder-Decoder on CAD classification problems such as cancer detection or metastasis grading.

Contribution

C.L. implemented the patch filtering, the training and evaluation of the Encoder-Decoder and GAN, performed hyperparameter tuning, and analyzed the results. S.M. implemented the Encoder-Decoder and Discriminator, reviewed literature and available data sets, implemented data pre-processing methods, and analyzed the results.

Ackowledgements

The authors would like to thank Professor Olivier Gevaert, Alexandre Momeni, and Marc Thibault for initiating the project idea, as well as Jeremy Irvin and Anand Avati for constructive advice on model selection and evaluation.

References

[1] P. A. Bautista, N. Hashimoto, and Y. Yagi, "Color standardization in whole slide imaging using a color calibration slide," *Journal of pathology informatics*, vol. 5, 2014.

[2] A. Sethi, L. Sha, A. R. Vahadane, R. J. Deaton, N. Kumar, V. Macias, and P. H. Gann, "Empirical comparison of color normalization methods for epithelial-stromal classification in h and e images," *Journal of pathology informatics*, vol. 7, 2016.

[3] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. de Souza, A. Baidoshvili, G. Litjens, B. van Ginneken, I. Nagtegaal, and J. van der Laak, "The importance of stain normalization in colorectal tissue classification with convolutional networks," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pp. 160–163, IEEE, 2017.

[4] 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.

[5] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[6] 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.

[7] P. Haub and T. Meckel, "A model based survey of colour deconvolution in diagnostic brightfield microscopy: Error estimation and spectral consideration," *Scientific reports*, vol. 5, p. 12096, 2015.

[8] H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using gan for histopathological images," *arXiv preprint arXiv:1710.08543*, 2017.

[9] A. Bentaieb and G. Hamarneh, "Adversarial stain transfer for histopathology image analysis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 792–802, 2018.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[11] G. Litjens, P. Bandi, B. E. Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels, *et al.*, "1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset," *GigaScience*, 2018.

[12] https://github.com/computationalpathologygroup/ASAP

[13] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision*, pp. 649–666, Springer, 2016.

[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

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

[19] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, p. 3, 2013.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[21] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 151–166, Springer, 2017.

[22] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. van der Laak, *et al.*, "Histopathology stain-color normalization using generative neural networks," 2018.

[23] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," *arXiv preprint arXiv:1804.01601*, 2018.

[24] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[28] P. Ahlgren, B. Jarneving, and R. Rousseau, "Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient," *Journal of the Association for Information Science and Technology*, vol. 54, no. 6, pp. 550–560, 2003.