Characterization of color normalization methods in digital pathology whole slide images

Ziaei, Dorsa, Li, Weizhe, Lam, Samuel, Cheng, Wei-Chung, Chen, Weijie

Dorsa Ziaei, Weizhe Li, Samuel Lam, Wei-Chung Cheng, Weijie Chen, "Characterization of color normalization methods in digital pathology whole slide images," Proc. SPIE 11320, Medical Imaging 2020: Digital Pathology, 1132017 (16 March 2020); doi: 10.1117/12.2550585

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Characterization of Color Normalization Methods in Digital Pathology Whole Slide Images

Dorsa Ziaei[1,2], Weizhe Li[2], Samuel Lam[2,3], Wei-Chung Cheng[2], Weijie Chen[2*]
[1]*University of Maryland Baltimore County (UMBC), Baltimore, MD, USA*
[2] *CDRH/OSEL/ DIDSR, Food and Drug Administration, Silver Spring, MD, USA*
[3]*University of Maryland, College Park, College Park, MD, USA*

ABSTRACT

The color rendering of whole-slide images (WSIs) depends on factors involving the sample, such as tissue type, preparation methods, staining type and staining protocol, as well as equipment, such as the WSI scanner, WSI viewer, and WSI display. Variations in any of these steps may change the color rendering and therefore affect the performance of pathologists in the interpretation of WSIs and the robustness of artificial intelligence algorithms. In the literature, color normalization techniques have been proposed to reduce the color variations. The purpose of this work is to develop an objective approach to characterizing color normalization methods used in digital pathology. We employed color normalization methods to normalize the color rendered by a WSI scanner and then compared the normalized color with the actual scan by that scanner. The normalization errors were evaluated on the pixel level using the CIE color difference ΔE metric that have been shown to correlate with visually perceived differences in human vision. A selected set of 310 patch images of breast tissues scanned by two scanners from the ICPR 2014 MITOS & ATYPIA contest was used. Images from one scanner were color normalized to match the color rendering of the other scanner. Four color normalization methods were compared – Macenko, Reinhard, Vahadane, and StainGAN. Experimental results show that average color differences between two scanners in terms of ΔE were reduced from 16.2 before normalization to the range of [13.7,16.9] after normalization for the Macenko, Reinhard, Vahadane methods, and to 8.3 for the StainGAN method. Apparently the StainGAN method is significantly superior to the other three methods in terms of the ΔE metric. As such, we demonstrated a quantitative method for objectively evaluating color normalization techniques. Future work is needed to explore the relationship of the color fidelity measure and the impact of color normalization on pathologist and AI performance in clinical tasks.

Keywords: whole slide image, digital pathology, color normalization

1. INTRODUCTION

Substantial color variations exist in digital pathology whole slide images (WSI) due to variations in tissue preparation, staining procedures, or digitization scanners (figure 1). Color variations may affect the performance of pathologists in the interpretation of WSIs and the robustness of Artificial Intelligence (AI) and Machine Learning (ML) algorithms applied to images from different sources. Color normalization techniques have been developed [1,2,3,4] to reduce color variations in WSIs prior to quantitative analysis with AI/ML algorithms. The theme of this paper is on characterization of such color normalization methods.

We begin with a review of some related work. Published studies on assessment of color normalization methods can generally be categorized into two types: (1) studies using a statistical image quality metric to quantify color difference between normalized images and target images; and (2) empirical studies comparing performance of AI/ML algorithms using a color normalization method and those without using one. Magee et al. [5] used Hotelling's T-square statistic to characterize the distributional difference of color channels between the normalized and the target images. This metric, however, is not intuitive to interpret. For the second type of studies, deep learning is a family of AI/ML algorithms that uses multiple artificial neural network layers to extract features from images for classification. Deep learning models such

* Send correspondence to: Weijie.chen@fda.hhs.gov; Phone: +1 301-796-2663.

Medical Imaging 2020: Digital Pathology, edited by John E. Tomaszewski, Aaron D. Ward, Proc. of SPIE Vol. 11320, 1132017 · © 2020 SPIE
CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2550585

as Convolutional Neural Networks (CNN) have been widely used for classification or segmentation of tissue specimens in histopathological applications [6]. Empirical studies have been reported in examining the effect of color normalization on CNN algorithms' classification accuracies with mixed or even contradictory results. Lee et al. [7] assessed Macenko [2], Khan [8] and Reinhard [1] color normalization methods and the study "did not show significant superiority in the CNN performance when color normalization was used to standardize the color appearance of histopathological image." Otálora et al. [9], on the other hand, showed that a color normalization method called domain adaptation in convolutional networks helped "model generalization to external datasets."

As such, while many color normalization techniques have been developed in the literature, consensus is lacking regarding methods for the assessment of these techniques. Our purpose in this work is to characterize color normalization techniques in WSI digital pathology using a color fidelity measure, namely the color difference ΔE defined by the International Commission on Illumination (CIE) standards. Our approach requires paired images, i.e., for a color normalization method intended to normalize images from two scanners, the assessment requires scanning the same slides with the two scanners. The ΔE is calculated at pixel level on paired images, thereby providing a more direct characterization of color difference.

In this paper, we studied four color normalization methods: the methods by Macenko [2], Reinhard [1], and Vahadane [3] and the StainGAN method [4]. While the first three methods require a single template image, the StainGAN approach used Generative Adversarial Network (GAN) to model color properties of two sets of training images. We developed a quantitative approach to characterize these color normalization methods, which may help make more informed decisions on choosing a color normalization method for training an AI/ML algorithm in digital pathology. In section 2, we introduce the publicly available dataset we used in this research. In section 3, we highlight the features and characteristics of each color normalization method. Section 4 describes color difference metric. In section 5 we show the results and in the last section, we discuss the results and conclud this paper.

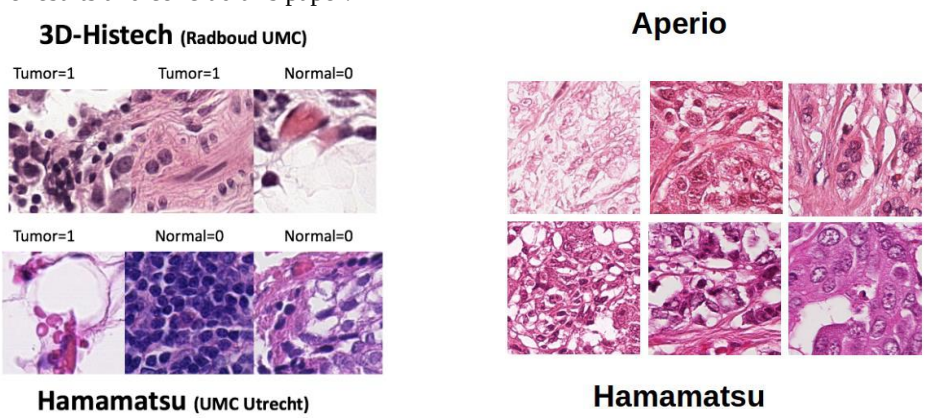

Figure. 1: Patch examples of WSIs from different scanners/medical centers
(left) CAMELYON dataset; (right) MITOS & ATYPIA dataset

2. DATASET

The dataset we use for demonstrating our assessment method is the MITOS & ATYPIA dataset for detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images [10]. The dataset was collected for a challenge at the International Conference on Pattern Recognition 2014 and has been made publicly available by the organizers. The slides were stained with standard Haemotoxylin and Eosin (H&E) dyes and were scanned by two scanners: Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT. It consists of 10x, 20x and 40x magnification levels selectable by pathologists.

We picked two patch images of size 256x256 pixels from the Hamamatsu scanner as the templates for the Macenko, Reinhard, and Vahadane methods. Each of these methods requires one template image to normalize the Aperio images. We repeated with two different templates to examine the variation of results. For StainGAN, we randomly extracted five patch images of 256x256 from each of 1045 40x frames, resulting in a total of 5225 image patches from each scanner. We

used 80% of these 5225x2 patches for training and 20% for internal validation. Because there are a total of 1200 40x frames in the dataset, we held out 155 independent frames for testing.

For the held-out dataset of 155 frames, we extracted two patch images from each frame for testing the color normalization methods. We observed apparent stitching errors in some frames and therefore extracted patch images of size 400x400 pixels to avoid the stitching boundaries, which led to a test set of 310 patches from each scanner.

2.1. Image Registration: Pixel-based color difference computation requires two images corresponding to the same slide scanned by two scanners be registered. We registered randomly selected 40x frame pairs using a feature-based method [11] implemented in the MATLAB Image Processing Toolbox. After visually checking the images and a trial-and-error procedure, we decided affine transformation is sufficient for registration in this study. Because our focus in this work is color normalization and registration error may limit our ability to measure the effect of color normalization, we selected patches that are well registered with an arbitrarily chosen threshold (0.93) on the registration accuracy (measured by correlation coefficient). Using this procedure, we selected 310 pairs of patch images.

3. STAIN COLOR NORMALIZATION METHODS

We investigated four color normalization methods reported in the literature including the methods of Reinhard [1], Macenko [2], Vahadane [3] and a method based on generative adversarial networks: the StainGAN method [4]. Below we introduce the algorithms they used for color normalization briefly.

3.1. Stain Color Matching

Reinhard et al. [1] proposed a color histogram matching based algorithm, based on linear transform in the perceptual LAB color space. In this approach, they matched means and standard deviations of each color channel in source and target domains. They segmented pixels into different classes and applied the linear transformation on each class of pixels separately. However, their approach raised new challenges such as segmentation of large images, which makes it difficult to classify all pixels. Additionally, Vahadane et al. showed that the statistical approach by Reinhard et al. [1] fills lumen with color when it should remain white after normalization.

3.2. Stain Color Extraction

Macenko et al. proposed stain color extraction in [2], which is an approach based on singular value decomposition (SVD) to decompose stain density maps for color normalization. Vahadane et al. showed that the Macenko et al. method may result in noisy images. Vahadane et al.[3] intended to extend and improve Macenko's technique by proposing a technique for stain separation and color normalization as a structure-preserving color normalization model. They applied non-negative matrix factorization to eliminate the negative component. They showed that their proposed algorithm performed qualitatively better in preserving structures, and in preserving stain densities, compared to the Macenko et al. and Reinhard et al. methods.

3.3. Image-to-Image Translation

A class of unsupervised Machine Learning based color normalization algorithms handles the problem of stain normalization as a style transfer problem and has achieved better results in terms of target image color and quantitative analysis. The goal of these models is to map color of the images in the source domain to the color of images in the target domain. The main properties of such algorithms are firstly to preserve the structure of the images after color normalization and secondly so that the color translation should be able to happen between any two images. Generative Adversarial Networks are the trending practical algorithms for normalizing stain color of histopathological images. Using GAN models, there is no need for template (reference) images anymore and color transfer can be done in both directions between two domains of images used for training the model.

We utilized the architecture of a generative algorithm, stainGAN. StainGAN is a cycleGAN-based [12] model, which was proposed by Shaban et al. and showed promising results on normalizing stained histopathological images. The GAN model consists of two networks, generator and discriminator, that are competing for generating fake image and discriminating the fake and real image. As a result, the generator network is able to generate fake images as close as possible to the real image. CycleGAN was designed to eliminate the need for paired training data by making a cycle in two domains of images as the source and target domains, and resulting in normalizing the color of source images to that of the target images and preserving the structure of the source images.

4. ΔE COLOR DIFFERENCE

We have a fully paired test dataset: 310 patch images from the Aperio scanner and 310 patch images from the Hamamatsu scanner, corresponding to the same slides. These images are registered to remove misalignment, allowing for a pixel-to-pixel comparison. We normalized the Aperio images to the color space of Hamamatsu images and used the acquired Hamamatsu as the ground truth. For each pixel, the color difference ΔE between two registered images of the same slide was calculated using the CIE color difference formula, a metric (ΔE) that correlates with visually perceived difference by human vision. Basically, ΔE calculates the Euclidean distance between two color coordinates in the CIELAB color space. Equation (1) shows how to calculate ΔE for each pixel in LAB color space:

$$\Delta E^*_{ab} = \sqrt{(L^*_2 - L^*_1)^2 + (a^*_2 - a^*_1)^2 + (b^*_2 - b^*_1)^2} \quad (1)$$

ΔE is a way of measuring the visible difference, or error between two colors mathematically [13]. It is a practical way to calculate the color similarity of two images. The LAB color space is used for calculating ΔE, since the LAB system's perceptually uniform basis makes these calculations more accurate than in other non-uniform systems such as RGB.

5. RESULTS

Figure 2 shows an example Aperio image normalized by four methods targeting colors of Hamamatsu images. The ΔE map indicates the change of color by the normalization method to the original image. The image scanned by the Hamamatsu scanner is shown on the right as a reference.

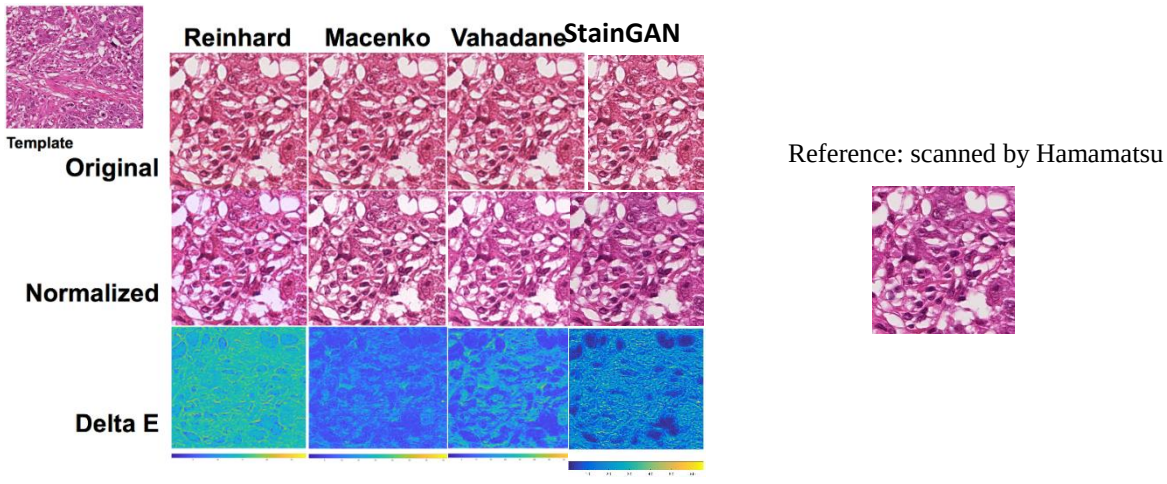

Figure 2. ΔE color difference between the normalized and original images. Top left corner: a template image from the Hamamatsu scanner used by the first three methods; top row: one image from the Aperio scanner; middle row: normalized image; bottom row: color difference between the original image and the normalized image. Right: reference image acquired from the Hamamatsu scanner.

Figure 3 shows histograms of average ΔE color difference between the Aperio and paired Hamamatsu images for both original Aperio (blue) and normalized Aperio (orange), over the 310 test image pairs. An effective normalization method is expected to shift the distribution to the left (i.e., reducing the color difference between the images scanned by two scanners). An ideal normalization method for perfectly registered images should yield a zero ΔE. The results indicate that both the Macenko and the Vahadane methods appear to be better than the Reinhard method. The StainGAN method apparently has the best ΔE performance among the four methods. We note that the StainGAN method not only resulted in lower ΔE values on average, but also the variability of the ΔE values among the images is lower. Furthermore, The Reinhard, Macenko, and Vahadane methods rely on a template image, which is another source of variability in these methods. Figure 4 shows the variations of these methods with respect to template images.

We further plot the histogram of the difference of ΔE: for each image, the difference of ΔE is calculated as *ΔE (between normalized Aperio and Hamamatsu)-ΔE (between original Aperio and Hamamatsu),* which is the reduction in ΔE due to normalization (figure 4). It is worth noting that the difference of ΔE for the StainGAN method is all below zero, meaning that for all the images in the test set the ΔE between the two scanners is reduced by the normalization method. For the other three methods, however, the color normalization sometimes increased the ΔE between the two scanners.

We statistically compared *ΔE (between original Aperio and Hamamatsu)* and *ΔE (between normalized Aperio and Hamamatsu)* using the paired *t* test on the 310 observations for each of the three normalization methods (Table 1). Consistent with observations from the histograms, the Macenko and the Vahadane methods appear to work similarly in terms of summary statistics whereas the Reinhard method works the worst. Again, the StainGAN method is shown to be better than the other three methods under comparison.

6. DISCUSSION AND CONCLUSIONS

We presented an objective approach to quantitatively characterize color normalization methods in digital pathology WSIs. The use of Euclidian distance ΔE in a device-independent color space (i.e., CIELAB) allowed quantified examination of color difference, which is a notion that formerly could only be described in subjective terms. The quantification of color fidelity was performed at the pixel level on paired and registered images, which allowed a direct characterization of the effect of color normalization by minimizing other confounding factors. Experimental results showed that average color differences between two scanners in terms of ΔE were reduced from 16.2 before normalization to the range of [13.7,16.9] after normalization for the Macenko, Reinhard, Vahadane methods, and to 8.3 for the StainGAN method. It appears that the StainGAN method is significantly better than the other three methods in terms of the ΔE metric.

The usefulness of color normalization is ultimately determined by its effect on the pathologist's diagnostic performance or its effect on AI/ML algorithm's performance in a clinical task. A question of interest is how the color fidelity metric is correlated with improvement of pathologist or algorithm performance. While color scientists suggest that a ΔE of 2.3 corresponds to a JND (just noticeable difference) [14], it remains an open question, to the best of our knowledge, how much color error is tolerated in pathologist's or algorithm's task performance. More fundamentally, we lack understanding of how color information versus morphological information is utilized by the pathologist or a deep learning algorithm in their diagnostic decision-making process. Future research in these areas is warranted.

In conclusion, our method is useful to objectively characterize color normalization techniques. We found that the StainGAN method performed better than the other three methods we compared in terms of the color fidelity metric. Our planned future work includes exploring the relationship of the color fidelity measure and the change in AI performance without relative to with color normalization.

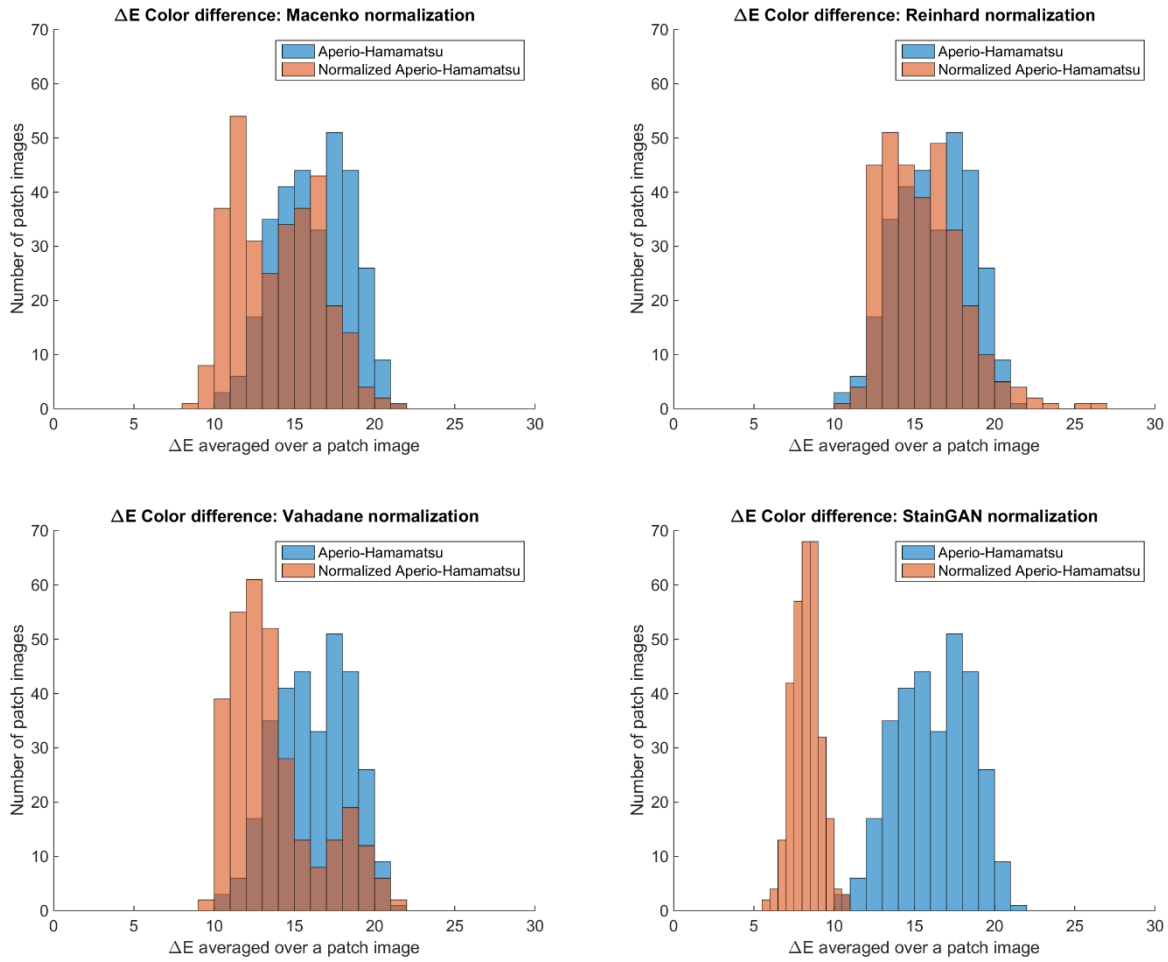

Figure 3. Histograms of ΔE color difference between the Aperio and paired Hamamatsu images for both unnormalized (blue) and normalized (orange) Aperio

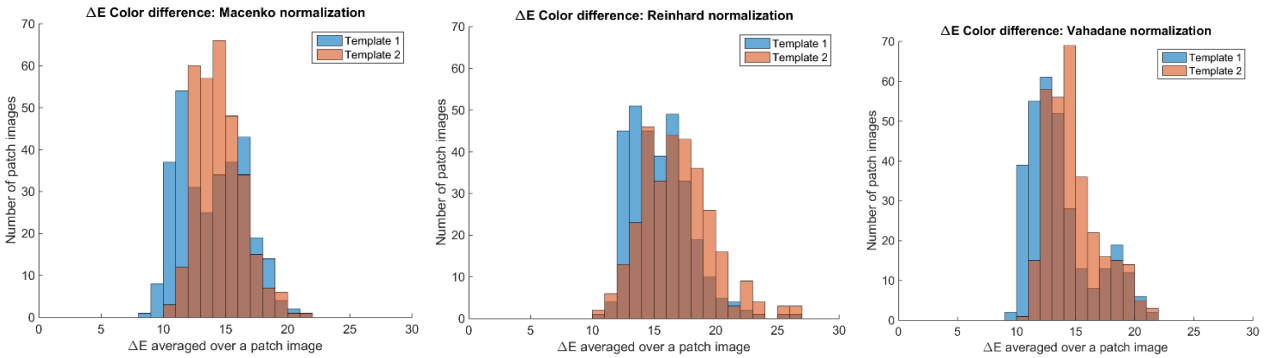

Figure 4. Variation of the Macenko, Reinhard, and Vahadane normalization methods with respect to template images

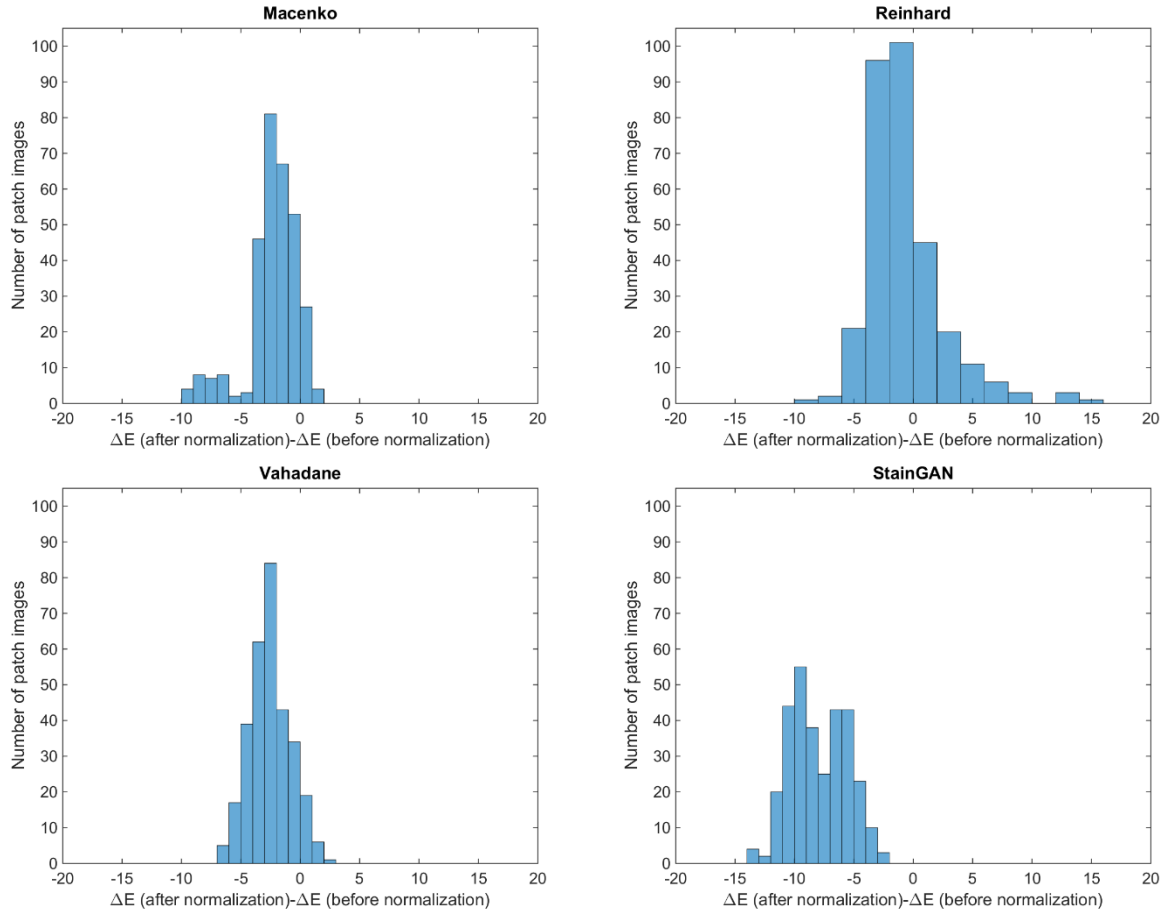

Figure 5: Histogram of the difference of ΔE (between Aperio and Hamamatsu) before and after normalization. Note that, for the StainGAN method, the ΔE difference is below zero for all the images.

Table 1: Statistical comparison of ΔE between the two scanners before color normalization vs. after color normalization

Template 1	**Average ΔE**	**Std(ΔE)**	**P value**	**CI of difference of ΔE**
No normalization	16.2	2.3		
Macenko	14.0	2.6	$< 10^{-4}$	[-2.5, -2.0]
Reinhard	15.5	2.5	$< 10^{-4}$	[-1.1, -0.4]
Vahadane	13.7	2.7	$< 10^{-4}$	[-2.7, -2.4]
Template 2	**Average ΔE**	**Std(ΔE)**	**P value**	**CI of difference of ΔE**
No normalization	16.2	2.3		
Macenko	14.5	1.9	$< 10^{-4}$	[-1.9, -1.5]
Reinhard	16.9	2.9	$< 10^{-4}$	[0.3, 1.0]
Vahadane	14.8	2.3	$< 10^{-4}$	[-1.6, -1.2]
StainGAN	8.3	0.9	$< 10^{-4}$	[-8.2, -7.7]

REFERENCES

[1] Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: "Color transfer between images". IEEE Computer Graphics and Applications, (2001)

[2] Macenko. M, et. al. "A method for normalizing histology slides for quantitative analysis". in Proc. IEEE Int. Symp. Biomed. Image, (2009)

[3] Vahadane. A, et. al. "Structure-preserving color normalization and sparse stain separation for histological images". IEEE transactions on medical imaging, 35(8):1961971, (Aug. 2016)

[4] Shaban, M. Tarek et al., "StainGAN: Stain Style Transfer for Digital Histological Images". IEEE 16th International Symposium on Biomedical Imaging, (2019)

[5] Magee, D., Treanor, D., Crellin, D., Shires, M., Smith, K., Mohee, K. and Quirke. P., "Color normalization in digital histopathology images". in Proc. Opt. Tissue Image Anal. Microsc., Histopathol. Endosc., pages 100–111, (2009)

[6] Sethi, A., Sha, L., Vahadane, A., Deaton, R.J., Kumar, N., Macias, V., & Gann, P.H. "Empirical comparison of color normalization methods for epithelial-stromal classification in H and E images". *Journal of pathology informatics, 7*, 17, (2016)

[7] Lee, G., Bajger, M., Clark, K., "Deep learning and color variability in breast cancer histopathological images: a preliminary study", 14th International Workshop on Breast Imaging (IWBI 2018)

[8] Khan, A.M., Rajpoot, N.M., Treanor, D., & Magee, D.R. A., "Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution". *IEEE Transactions on Biomedical Engineering, 61*, 1729-1738, (2014)

[9] Otálora, S., Atzori, M., Andrearczyk, V., Khan, A., & Müller, H., "Staining Invariant Features for Improving Generalization of Deep Convolutional Neural Networks in Computational Pathology". *Front. Bioeng. Biotechnol*, (2019)

[10] https://mitos-atypia-14.grand-challenge.org/

[11] Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool. "SURF: Speeded Up Robust Features." *Computer Vision and Image Understanding (CVIU)*.Vol. 110, No. 3, pp. 346–359, (2008)

[12] Zhu, J., Park, T., Isola, P., & Efros, A.A. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". *IEEE International Conference on Computer Vision (ICCV)*, 2242-2251, (2017)

[13] Sharma, G., Wu, W., & Dalal, E.N. "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations", (2005)

[14] Sharma, Gaurav *Digital Color Imaging Handbook (1.7.2 ed.). CRC Press. ISBN 0-8493-0900-X,* (2003)