

StainGAN: Stain Style Transfer for Digital Histological Images

M Tarek Shaban¹, Christoph Baur¹, Nassir Navab^{1,2}, and Shadi Albarqouni¹

¹ Computer Aided Medical Procedures (CAMP), Technische Universität München, Munich, Germany

² Whiting School of Engineering, Johns Hopkins University, Baltimore, USA
first.last@tum.de

Abstract. Digitized Histological diagnosis is in increasing demand. However, color variations due to various factors are imposing obstacles to the diagnosis process. The problem of stain color variations is a well-defined problem with many proposed solutions. Most of these solutions are highly dependent on a reference template slide. We propose a deep-learning solution inspired by CycleGANs that is trained end-to-end, eliminating the need for an expert to pick a representative reference slide. Our approach showed superior results quantitatively and qualitatively against the state of the art methods (10% improvement visually using SSIM). We further validated our method on a clinical use-case, namely Breast Cancer tumor classification, showing 12% increase in AUC. Code will be made publicly available.

Keywords: Histology Images, Stain Normalization, Generative Adversarial Networks, Deep Learning.

1 Introduction

Histology relies on the study of microscopic images to diagnose disease based on the cell structures and arrangements. Staining is a crucial part of the tissue preparation process. The addition of staining components (mainly Hematoxylin and Eosin) transforms the naturally transparent tissue elements to become more distinguishable (hematoxylin dyes the nuclei a dark purple color and the eosin dyes other structures a pink color). However, results from the staining process are inconsistent and prone to variability; due to differences in raw materials, staining protocols across different pathology labs, inter-patient variabilities, and slide scanner variations as shown in 1. These variations not only cause inconsistencies within pathologists [7] but it also hinders the performance of Computer-Aided Diagnosis (CAD) systems [5].

Stain normalization algorithms have been introduced to overcome this well-defined problem of stain color variations. These methods can be broadly classified into three classes, **Color matching based methods** that try to match the color spectrum of the image to that of the reference template image. Reinhard et al.[11] align the color-channels to match that of the reference image in the

lab color model, However, this can lead to improper color mapping, as the same transformation is applied across the images and does not take into account the independent contribution of stain dyes to the final color.

Further, there are **Stain-separation methods** where normalization is done on each staining channel independently. For instance, Macenko et al. [10] find the stain vectors by transforming the RGB to the Optical Density (OD) space. On the other hand, the method proposed by Khan et al. [9] estimate the stain matrix based on a color-based classifier that assigns every pixel to the appropriate stain component. According to Babak et al [2], these methods do not take the spatial features of the tissue structure into account, which leads to improper staining. Nevertheless, most of the methods in this class rely on an expertly picked reference template image, which has a major effect on the outcome of the methods as we show later. The third class of methods is the **Pure learning based approaches**, that handle the problem of stain normalization as a style-transfer problem, BenTaieb et al.[3] handle the problem using auxiliary classifier GAN with an auxiliary task on top (classifier or segmentation). Our method (StainGAN) is also GAN based but as opposed to BenTaieb does not require to be trained for a specific task in order to achieve stain style transfer.

We propose a method based on Generative adversarial networks (GANs) that not only eliminates the need for the reference image but also achieves high visual similarity to the target domain, making it easier to get rid of the stain variations thus improving the diagnosis process for both the pathologist and CAD systems. StainGAN is based on the Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks (CycleGAN)[16]. Cycle-consistency allows the images to be mapped to different color-model but preserves the same tissue structure. We evaluate our approach against the state of the art methods quantitatively and indirectly through a breast tumor classification. Our contributions include 1) Using Style-transfer for the classic stain normalization problem and achieving better results quantitatively than the state of the art methods. 2) Removing the need for a manually picked reference template, as our model learn the whole distribution. 3) Providing a benchmark for comprehensive comparison between the proposed StainGAN method and most of the state of the art methods.

2 Methodology

Our framework, as depicted in Fig. 2, employs the CycleGAN concept [16] to transfer the H&E Stain Appearance between different scanners, *i.e* from Hamamatsu (H) to Aperio (A) Scanner, without the need of paired data from both domains. The model consists of two generator and discriminator pairs, the first pair (G_H and D_H), tries to map images from domain A to domain H $G_H : \mathcal{X}_A \rightarrow \mathcal{X}_H$. While the Generator G_H tries to generate images that match domain H , the discriminator D_H tries to verify if images come from the real domain H or the fake generated ones. The other pair (G_A and D_A), undergoes the same process in the

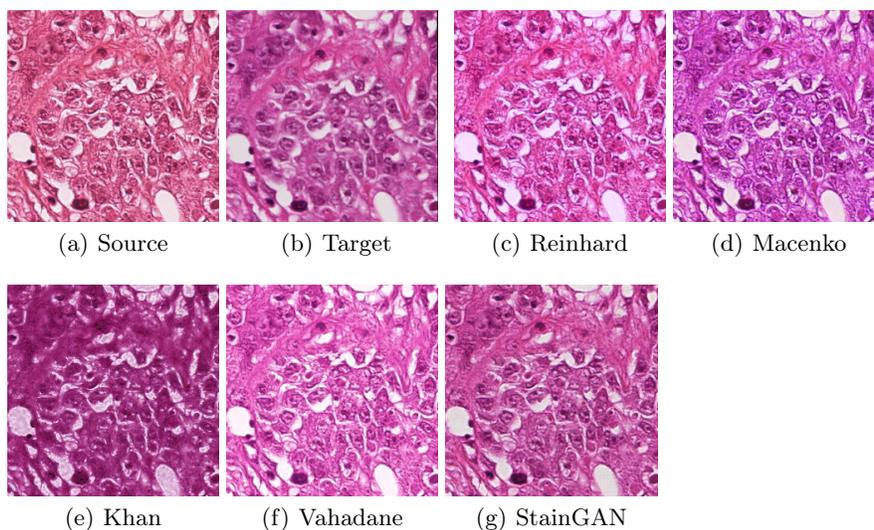


Fig. 1: Stain Normalization of various methods, the goal is to match the Target image.

reverse direction, $G_A : \mathcal{X}_H \rightarrow \mathcal{X}_A$, as

$$\hat{X}_H = G_H(X_A; \theta_H), \quad \hat{X}_A = G_A(\hat{X}_H; \theta_A), \quad \text{s.t. } d(X_A, \hat{X}_A) \leq \epsilon, \quad (1)$$

and

$$\hat{X}_A = G_A(X_H; \theta_A), \quad \hat{X}_H = G_H(\hat{X}_A; \theta_H), \quad \text{s.t. } d(X_H, \hat{X}_H) \leq \epsilon, \quad (2)$$

where $d(\cdot, \cdot)$ is a distance metric between the given image and the reconstructed one, so-called a *Cycle-Consistency* constraint, and both θ_A and θ_H are the model parameters.

To achieve this, the models are trained to meet the following objective function,

$$\mathcal{L} = \mathcal{L}_{Adv} + \lambda \mathcal{L}_{Cycle}, \quad (3)$$

where \mathcal{L}_{Adv} is the adversarial loss, \mathcal{L}_{Cycle} is the cycle-consistency loss, and λ is a regularization parameter.

The adversarial loss tries to match the distribution of the generated images to that of the target domain (*Forward Cycle*), and match the distribution of the generated target domain back to the source domain (*Backward Cycle*) as $\mathcal{L}_{Adv} = \mathcal{L}_{GAN}^A + \mathcal{L}_{GAN}^H$, where \mathcal{L}_{GAN}^A is given as

$$\begin{aligned} \mathcal{L}_{GAN}^A(G_A, D_A, X_H, X_A) = & \mathbb{E}_{X_A \sim p_{data}(X_A)} [\log D_A(X_A)] \\ & + \mathbb{E}_{X_H \sim p_{data}(X_H)} [\log(1 - D_A(G_A(X_H; \theta_A)))] \end{aligned} \quad (4)$$

The Cycle-consistency loss ensures that generated images preserve similar structure as in the source domain. This loss goes in both directions forward and

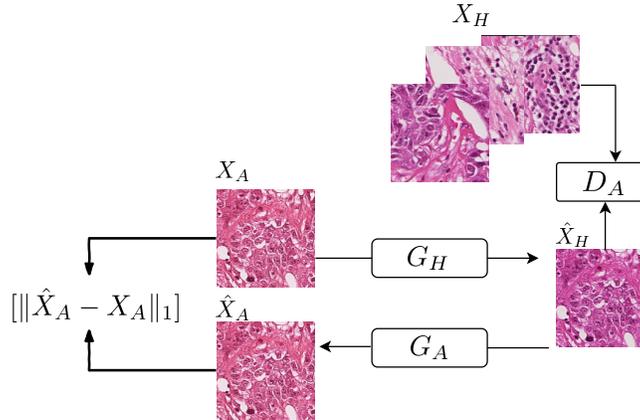


Fig. 2: Our StainGAN framework: Images are mapped to domain H and then back to domain A to ensure structure perseverance. Same process is done in reverse direction from domain H to domain A.

backward cycles to assure stability.

$$\begin{aligned} \mathcal{L}_{Cycle}(G_A, G_H, X_A, X_H) = & \mathbb{E}_{X_A \sim p_{data}(X_A)} [\|G_A(G_H(X_A; \theta_H); \theta_A) - X_A\|_1] \\ & + \mathbb{E}_{X_H \sim p_{data}(X_H)} [\|G_H(G_A(X_H; \theta_A); \theta_H) - X_H\|_1], \end{aligned} \quad (5)$$

where $\|\cdot\|_1$ is the ℓ_1 -norm.

Network Architectures. In our model, ResNet [6] and 70×70 PatchGAN [8] are employed as network architecture for the generators and the discriminators, respectively.

3 Experiments and Results

To have a fair comprehensive comparison, we evaluate our model against the state of the art methods of Reinhard[11] Macenko[10], Khan[9], and Vahadane[13] as follows; i) *Quantitative comparison* between the visual appearances of stained images, and the effect of varying reference slides, ii) *Quantitative comparison* of the Hematoxylin and Eosin stain vectors separation, and iii) *Use-case experiment*, stain normalization as a preprocessing step in the pipeline of tumor classifier.

3.1 Stain Transfer

The goal is to be able to map the patches from scanner A (Aperio) to that of Scanner H (Hamamatsu), the dataset includes same tissue sections scanned with

both scanners, Slides from scanner A are normalized to match scanner H, then compared with the real slides of Scanner H (ground truth). Results are evaluated using various similarity matrices. Additionally, Stain Vectors are extracted using Ruifrok’s method [12] and compared to that of the ground truth. Visual results of the of the various methods are illustrated in 1. It is clear that our results are more close visually to the target images.

Dataset. The dataset is publicly available as part of the MITOS-ATYPIA 14 challenge ³. Dataset consists of 284 frames at $X20$ magnification which are stained with standard Hematoxylin and Eosin (H&E) dyes. Same tissue section has been scanned by two slide scanners: Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT. Slides from both scanners resized to have equal dimensions of (1539×1376) . Rigid registration was performed to eliminate any misalignment. For the training set, we extract 10,000 random patches from the first 184 full slide of both scanners. For evaluation, 500 same view section patches of 256×256 were generated from each of the two scanners from the last 100 full slides (unseen in the training set), patches from the Scanner H are used as the ground truth.

Implementation. For the state of the art methods, a reference slide was picked carefully. Our method does not require a reference slide as it learns the distribution of the whole data, model was trained using 10,000 random unpaired patches from both scanners for 26 epochs with the regularization parameter set to $\lambda=10$, learning rate is set to 0.0002, Adam optimizer with a batch size of 4 is used. Hardware of GeForce GTX TITAN X 12GB and Pytorch framework were used.

Evaluation Metrics. Results are compared to the ground truth using four similarity measures: Structural Similarity index (SSIM)[14], Feature Similarity Index for Image Quality Assessment (FSIM) [15], Peak Signal-to-Noise Ratio (PSNR) and Pearson correlation coefficient similarity[1]. On the other hand, Euclidian norm distance as reported in [2] was used to evaluate stain vectors against the ground truth.

Results. As reported in Table 1, our results significantly outperform the state-of-the-art methods in all similarity metrics ($p < 0.01$). Further, it has shown a better stain seperability compared to the ground-truth stain vectors as shown in Table. 2.

Slide Reference Sensitivity. We ran the same experiment as in the previous one, but with three different reference images shown in changes of the SSIM score with respect to the reference images, results are reported in Fig.3, which shows that the conventional methods are sensitive to the reference image.

³ <https://mitos-atypia-14.grand-challenge.org>

Table 1: Stain Transfer Comparison: Mean \pm Standard Deviation

Methods	SSIM	FSSIM	Pearson Correlation	PSNR
Reinhard [11]	0.58 ± 0.10	0.67 ± 0.05	0.51 ± 0.21	13.4 ± 1.61
Macenko [10]	0.56 ± 0.12	0.67 ± 0.05	0.45 ± 0.22	14.0 ± 1.68
Khan [9]	0.67 ± 0.11	0.71 ± 0.05	0.54 ± 0.20	16.3 ± 2.11
Vahadane [13]	0.65 ± 0.13	0.71 ± 0.06	0.53 ± 0.22	14.2 ± 2.13
StainGAN	0.71 ± 0.11	0.73 ± 0.06	0.56 ± 0.22	17.1 ± 2.50

Table 2: Stain Vectors Comparison: Mean \pm Standard Deviation (the lower the better) and Processing time taken to normalize the 500 images

Methods	Staining Separation		Processing time
	S_H	S_E	Time (sec)
Reinhard [11]	20.7 ± 4.85	18.0 ± 3.79	9.80
Macenko [10]	21.05 ± 4.11	12.2 ± 3.14	63.63
Khan [9]	21.2 ± 3.65	12.0 ± 2.73	2196.25
Vahadane [13]	20.0 ± 4.57	12.9 ± 3.45	582.94
StainGAN	18.0 ± 4.84	10.0 ± 3.32	74.55

3.2 Use-Case Application

Stain-normalization is a mandatory preprocessing step in most CAD systems and has proven to increase performance[5]. The aim of this experiment is to show the performance of stain normalization in the context of breast cancer classification, model is trained with patches from the lab 1, testing is done with patches from lab 2 (different staining appearances). We normalize the test-set to match lab 1 and compare results accordingly.

Dataset. Camelyon16 challenge ⁴ consisting of 400 whole-slide images collected in two different labs in Radboud University Medical Center (lab 1) and University Medical Center Utrecht (lab 2). Otsu thresholding was used to remove the background, Afterwards, 40,000 256×256 patches were generated on the x40 magnification level, 30,000 were used for training and 10,000 used for validation from lab 1 and 10,000 patches were generated for testing from lab 2.

Implementation. The classifier network is composed of three convolutional layers with a ReLU activation, followed by a max-pooling layer turning the given image into a logit vector. Classifier has been trained for 30 epochs with RMSprop optimizer and binary cross-entropy loss. For the stain normalization, representative reference slide from lab 2 was picked for the conventional methods, Our method was trained using the same parameters from the previous experiment on training set: 68000 patches from both labs.

⁴ <https://camelyon16.grand-challenge.org/>

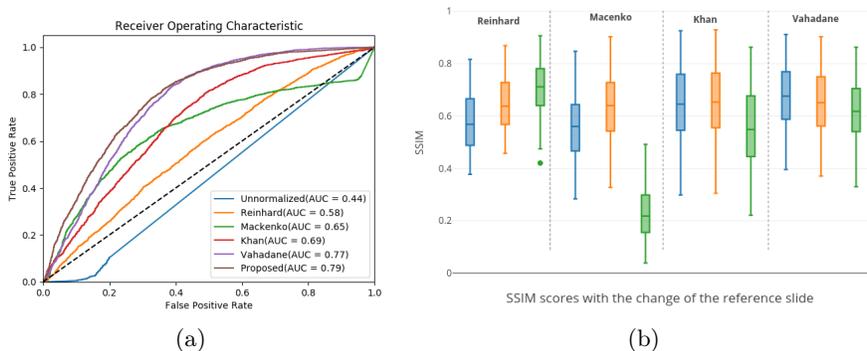


Fig. 3: (a) ROC curves of the test-set pre-processed using different stain normalization methods. (b) Box plot shows the variation of the SSIM metric due to the improper selection of reference slide.

Evaluation Metrics. To evaluate the classifier performance, we report the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC).

Results. The ROC curves and AUC of different classifiers, trained on different stain normalization methods, are presented in Fig.3. Our proposed StainGAN method shows a relative improvement of (80%, 36%, 22%, 15%, and 3%) on AUC over the Un-normalized, Reinhard, Macenko, Khan, and Vahadane, respectively.

4 Discussion and Conclusion

In the paper, we presented StainGAN as a novel method for the stain normalization task. Our experiments revealed that our method significantly outperforms the state of the art. The visual appearance of different methods can be seen in Fig.1. It clearly shows that images normalized with StainGAN are very similar to the ground truth. Further, our StainGAN method has been validated in a clinical use-case, namely Tumor Classification, as a pre-processing step showing a superior performance. Moreover, the processing time of our method is on par with Macenko as reported in Table. 2. We believe that end-to-end learning based approaches are ought to overtake the classic stain normalization methods. Yet, there is still more room for improvement, for example, we can think of Unified representation, similar to [4], that can map many to many stain style domains.

Acknowledgment

This work was partially funded by the German Research Foundation (DFG, SFB 824), and Bavarian Research Foundation (BFS, IPN2).

References

1. Ahlgren, P., Jarneving, B., Rousseau, R.: Requirements for a cocitation similarity measure, with special reference to pearson’s correlation coefficient. *Journal of the Association for Information Science and Technology* 54(6), 550–560 (2003)
2. Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., van der Laak, J.A.: Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging* 35(2), 404–415 (2016)
3. BenTaieb, A., Hamarneh, G.: Adversarial stain transfer for histopathology image analysis. *IEEE Transactions on Medical Imaging* (2017)
4. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020* (2017)
5. Ciompi, F., Geessink, O., Bejnordi, B.E., de Souza, G.S., Baidoshvili, A., Litjens, G., van Ginneken, B., Nagtegaal, I., van der Laak, J.: The importance of stain normalization in colorectal tissue classification with convolutional networks. *arXiv preprint arXiv:1702.05931* (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
7. Ismail, S.M., Colclough, A.B., Dinnen, J.S., Eakins, D., Evans, D., Gradwell, E., O’sullivan, J.P., Summerell, J.M., Newcombe, R.G.: Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *Bmj* 298(6675), 707–710 (1989)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. *arXiv preprint* (2017)
9. Khan, A.M., Rajpoot, N., Treanor, D., Magee, D.: A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering* 61(6), 1729–1738 (2014)
10. Macenko, M., Niethammer, M., Marron, J., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: *Biomedical Imaging: From Nano to Macro, 2009. ISBI’09. IEEE International Symposium on*. pp. 1107–1110. IEEE (2009)
11. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. *IEEE Computer graphics and applications* 21(5), 34–41 (2001)
12. Ruifrok, A.C., Johnston, D.A., et al.: Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology* 23(4), 291–299 (2001)
13. Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N.: Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging* 35(8), 1962–1971 (2016)
14. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4), 600–612 (2004)
15. Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* 20(8), 2378–2386 (2011)
16. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* (2017)