StyPath: Style-Transfer Data Augmentation for Robust Histology Image Classification

Pietro Antonio Cicalese[1,2(✉)], Aryan Mobiny[1], Pengyu Yuan[1], Jan Becker[3], Chandra Mohan[2], and Hien Van Nguyen[1]

[1] Department of Electrical Engineering, University of Houston, Houston, TX, USA
[2] Department of Biomedical Engineering, University of Houston, Houston, TX, USA
pcicalese@uh.edu
[3] Institute of Pathology, University Hospital of Cologne, Cologne, Germany

Abstract. The classification of Antibody Mediated Rejection (AMR) in kidney transplant remains challenging even for experienced nephropathologists; this is partly because histological tissue stain analysis is often characterized by low inter-observer agreement and poor reproducibility. One of the implicated causes for inter-observer disagreement is the variability of tissue stain quality between (and within) pathology labs, coupled with the gradual fading of archival sections. Variations in stain colors and intensities can make tissue evaluation difficult for pathologists, ultimately affecting their ability to describe relevant morphological features. Being able to accurately predict the AMR status based on kidney histology images is crucial for improving patient treatment and care. We propose a novel pipeline to build robust deep neural networks for AMR classification based on StyPath, a histological data augmentation technique that leverages a light weight style-transfer algorithm as a means to reduce sample-specific bias. Each image was generated in 1.84 ± 0.03 s using a single GTX TITAN V gpu and pytorch, making it faster than other popular histological data augmentation techniques. We evaluated our model using a Monte Carlo (MC) estimate of Bayesian performance and generate an epistemic measure of uncertainty to compare both the baseline and StyPath augmented models. We also generated Grad-CAM representations of the results which were assessed by an experienced nephropathologist; we used this qualitative analysis to elucidate on the assumptions being made by each model. Our results imply that our style-transfer augmentation technique improves histological classification performance (reducing error from 14.8% to 11.5%) and generalization ability.

Keywords: Pathology · Style-transfer · CNN classifier · Data augmentation · Inter-observer agreement

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-59722-1_34) contains supplementary material, which is available to authorized users.

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12265, pp. 351–361, 2020.
https://doi.org/10.1007/978-3-030-59722-1_34

1 Introduction

Low intra and inter observer reproducibility in the analysis of histologic features has been a persistent concern amongst pathologists across various tissue types and diseases. Although concordance is generally within an acceptable range for diagnosis, the clinical adoption of certain scoring parameters has been hindered by disagreement between observers and insufficient empirical evidence. This has a negative impact on the reproducibility of any given histological diagnosis and thus complicates the path to both consistent and effective treatment, and the design of multicentric pharmaceutical studies. The degree of disagreement observed has been the topic of several publications; varying levels of experience, the number of scoring parameters, cutoff values, and the type of stains used to assess a biopsy have all been associated with variable diagnosis [6,13,15,23]. Discrepancies in staining quality within and between pathology labs have also been implicated as a cause for variability, with variations in stain intensity being associated with poor diagnostic performance [2]. The effect of this apparent stain bias is similar to the texture and context bias we observe when training Convolutional Neural Networks (or CNNs), which negatively impacts the generalization ability of the model.

Geirhos *et al.* showed that CNNs trained on the ImageNet dataset are heavily biased towards texture and fail to match human performance in the silhouette and edge classification tasks [10]. These models can therefore appear to yield exceptional performance in real world applications, when they are actually depending on a single cue for classification. This poses a significant challenge in histological diagnosis tasks where each biopsy has unique texture and color characteristics, which may have a negative impact on performance and the model's ability to generalize. This issue is also exacerbated by the relatively small sample sizes of histology datasets; whereas large datasets are less likely to yield models with sample specific bias, a small dataset is likely to yield models with poor generalization ability. We believe that by combining texture and color information from various training samples, we could reduce this texture bias while increasing the classifier shape bias, ultimately yielding a more robust model. Our contributions are as follows:

- We propose a sample and condition agnostic style-transfer data augmentation technique to improve histological training performance with a small glomerular-level Antibody Mediated Rejection (AMR) dataset. Our technique depends on both the moderated transfer of spatial relationships and style, yielding a more diverse training population that retains the desired target concepts.
- We use Bayesian inference to compensate for the ignorance of the model (due to the small dataset size) and to estimate a more reliable measure of performance. Our technique improves upon the prediction accuracy of the baseline model at a lower computational cost relative to similar techniques.
- Qualitative analysis and visual inspection by an experienced nephropathologist are provided to demonstrate that the StyPath augmented models are

more robust, capturing information that would have otherwise been missed by the baseline model.

Related Works. The issue of histological stain normalization has been addressed by various research groups in both the medical and computer science communities. Adam *et al.* proposed a single feature schema in the qualitative scoring of polyomavirus BK (BKV) nephropathy immunohistochemistry (IHC) slides and stain protocol standardisation practices as a means to reduce the inter-observer variance associated with differing stain intensities [2]. This solution is limited by the difficulty associated with novel scoring parameter adoption and the cost incurred by pathology labs to standardize their equipment and procedures. These kinds of limitations have ultimately led various groups to develop *in silico* stain normalization techniques that could be used to improve upon quantitative histological image analysis. In the hemotoxylin and eosin (H&E) data normalization task, Macenko *et al.* proposed a deconvolution technique which effectively separates the hemotoxylin and eosin stains to then generate their respective normalized components [14]. Bejnordi *et al.* developed a whole-slide image color standardizer that combines color and spatial information to categorize each pixel in a slide scan as a stain component [3]. Bug *et al.* proposed a style-transfer normalization algorithm which transfers pathology lab profiles between samples without altering spatial relationships [5]. Interestingly, Tellez *et al.* evaluated all three of these stain normalization techniques and showed that they had a largely negligible (or even negative) effect on CNN classification performance while adding a significant computational cost, prompting them to favor other stain augmentation techniques [22]. It is important to note that the style-transfer algorithm they evaluated only transfers the color profiles between pathology labs, remaining largely agnostic to the sources of variability present within each lab and stained section. Shaban *et al.* and BenTaieb *et al.* also proposed conditional stain transfer techniques that utilized adversarial networks to generate new samples for each target domain [4,20]. While powerful, these techniques require explicitly defined domain labels; this prevents the techniques from capturing important variations that may be present within each predefined condition set. We were interested in developing a sample and condition agnostic style-transfer augmentation procedure that could be applied to any histological dataset to address all sources of stain variability. We also propose the moderated transfer of spatial relationships between samples; while prior techniques avoided altering the morphological appearance of each image, we hypothesized that slight image alterations through the style-transfer algorithm could be beneficial to the generalization capabilities of a model, especially when trained with a small dataset.

2 Methodology

AMR Dataset Generation and Annotation. A total of 86 (38 non-AMR and 48 AMR, chronic active AMR, or chronic AMR) blood group ABO- compatible kidney transplant biopsies were randomly selected for processing and

analysis, each fulfilling the minimum sample criteria ($\geq$7 glomeruli, $\geq$1 artery). Each paraffin embedded section was cut at 2 mm and stained with periodic acid-Schiff (PAS) in the same pathology lab over the span of two years. For each section, micrographs were taken from all non-globally sclerosed glomeruli that were at least four levels apart at a resolution of 1024×768 pixels. All segmented glomerular images were then annotated by an experienced nephropathologist using the Labelbox platform; label choices were given as either AMR, non-AMR, or inconclusive, yielding 1503 conclusively labeled glomerular images (1001 non-AMR and 502 AMR) [1].

Style-Transfer. One can think of the style-transfer algorithm as a means to generate artificial images that combine the high-level semantic representation of one image with the low-level perceptual representation of another image. The semantic image (i.e. the content image) represents the objects that will be depicted in the generated image, while the perceptual image (i.e. the style image) characterizes simpler information (such as color and texture). The ability to generate these artificial images depends on the extraction and manipulation of feature maps from the filters of the selected layers of a CNN [9]. An input image $\mathbf{x}$ is effectively encoded in each layer of a CNN by the filter responses to image $\mathbf{x}$, which allows us to extract various representations of the image depending on the layer. Feature maps from deeper layers of a CNN characterize complex concepts (such as cells and glomeruli) while those from shallow layers characterize simple concepts (such as edges and color). We say that a layer l with N_l unique filters therefore has N_l unique feature maps of dimensions M_l^{hw}, corresponding to the height multiplied by the width of each feature map. We can then say that the filter responses within a given layer l can be stored in a matrix $F_l \in R^{N_l \times M_l^{hw}}$ where F_l^{ij} corresponds to the activation of the i^{th} filter at position j in layer l.

Suppose that we have a given content image $\mathbf{x}_{cont}$ and an output image $\mathbf{x}_{out}$; we define their respective feature responses in a given layer l as F_l^{cont} and F_l^{out}, respectively. We can then define the squared-error content loss λ^{cont} for a given layer l following

$$\lambda_l^{cont}(\mathbf{x}_{cont}, \mathbf{x}_{out}) = \frac{1}{2} \sum_{ij} (F_l^{ij,out} - F_l^{ij,cont})^2. \tag{1}$$

We can thus define the derivative of the squared-error loss between each set of feature representations for a given layer l following

$$\frac{\mathrm{d}\lambda_l^{cont}}{\mathrm{d}F_l^{ij,out}} = \begin{cases} (F_l^{ij,out} - F_l^{ij,cont}), & F_l^{ij,out} > 0 \\ 0, & F_l^{ij,out} < 0 \end{cases}, \tag{2}$$

which can then be used to compute the content gradient with respect to image $\mathbf{x}_{out}$. We then generate a Gram matrix which takes the correlation between various filter responses as a means to capture the style of the desired style image $\mathbf{x}_{sty}$. The generated Gram matrix is given by $G_l \in R^{N_l \times M_l^h \times M_l^w}$, where G_l^{ij}

represents the inner product between the vectorised feature map i and j in a given layer l at position k. We define G_l^{ij} following

$$G_l^{ij} = \sum_k F_l^{ik} F_l^{jk}, \tag{3}$$

which we then use to match the style between $\mathbf{x}_{out}$ and $\mathbf{x}_{sty}$ by minimising the mean-squared distance between their respective Gram matrices. Let $G_l^{ij,out}$ and $G_l^{ij,sty}$ represent the Gram matrices from a given layer l of $\mathbf{x}_{out}$ and $\mathbf{x}_{sty}$, respectively. We then say that the contribution of layer l to the total style loss is given by

$$\lambda_l^{sty} = \frac{1}{4(N_l)^2 (M_l^{hw})^2} \sum_{ij} (G_l^{ij,out} - G_l^{ij,sty})^2, \tag{4}$$

while total style loss is given by

$$\lambda^{sty}(\mathbf{x}_{sty}, \mathbf{x}_{out}) = \sum_{l=0}^{L} w_l \lambda_l^{sty}, \tag{5}$$

with w_l corresponding to weighting factors of the contribution of each layer to the total loss (which we simply set equal to one divided by the number of active layers). We can then compute the derivative of λ_l^{sty} following

$$\frac{\mathrm{d}\lambda_l^{sty}}{\mathrm{d}F_l^{ij}} = \begin{cases} \frac{1}{(N_l)^2 (M_l^{hw})^2} (F_l^{ij})^T (G_l^{ij,out} - G_l^{ij,sty}) & F_l^{ij} > 0 \\ 0 & F_l^{ij} < 0 \end{cases}, \tag{6}$$

thus allowing us to compute the style gradient with respect to $\mathbf{x}_{out}$. Finally, to generate the style-transfer samples, we simply minimize both λ^{cont} and λ^{sty} following

$$\lambda^{tot}(\mathbf{x}_{cont}, \mathbf{x}_{sty}, \mathbf{x}_{out}) = \alpha \lambda^{cont}(\mathbf{x}_{cont}, \mathbf{x}_{out}) + \lambda^{sty}(\mathbf{x}_{sty}, \mathbf{x}_{out}), \tag{7}$$

where α is used to scale down the content loss, allowing us to control the stylization of the generated image.

Approximate Bayesian Inference via MC-Dropout. Training a standard neural network parameterized by its weights is equivalent to generating a maximum likelihood estimation (MLE) of the network parameters, which yields a single set of parameters [11]. Such a model generates point estimates for each testing sample it classifies and ignores any model uncertainty that may be present in the proper weight values. In medical applications, this can eventually mislead the physician into believing that a model is confident about a prediction that may actually be a lucky guess [7]. Model uncertainty, also known as epistemic uncertainty, is most prevalent when a model is trained using a small sample set, where irrelevant information may be abused by the model to improve performance [7,16]. It would therefore be more informative to generate a model that

provides a probabilistic estimate of its predictions, which can then be used to estimate the model's level of uncertainty. This can be accomplished by generating a Bayesian Neural Network (BNN) model; it is possible to generate a prior distribution over the network's parameters, outputting a probability distribution which can be used to estimate class posterior probabilities for each testing sample [18]. One can then integrate over the class posterior probabilities to produce a predictive posterior distribution over the class membership probabilities, and measure dispersion over the predictive posterior to generate uncertainty estimates. However, BNN models are computationally intractable, which has prompted various groups to develop methods to approximate Bayesian inference. To generate such a model, Gal *et al.* proved that a feed-forward neural network with a given number of layers and non-linearities can be equivalent to approximate variational inference in the deep Gaussian Process model when dropout is applied to all units [8]. We can therefore use Monte Carlo (MC) dropout at test time to yield a Bernoulli distribution over the weights of a CNN to generate an approximation of the posterior distribution without having to train additional parameters [17]. Through this technique, we can therefore quantitatively measure a more accurate estimate of each model's performance and describe their ability to generalize during the testing phase.

Grad-CAM Visualization. Being able to visualize and interpret the classification criteria being used by a CNN is critical to both confirming the assumptions being made by the classifier and learning from the classifiers decisions. Selvaraju *et al.* showed that visual explanations for each model prediction could be generated by using gradient information flowing into the last convolutional layer of a CNN, capturing the semantic information being used to classify a given sample [19]. These extracted Gradient-weighted Class Activation Maps (Grad-CAMs) could thus be used to depict a high-level representation of the classifiers decision making process. To further understand how both the baseline and Sty-Path augmented models drew their conclusions, we chose to generate Grad-CAM representations of each testing set prediction, which were then assessed by an experienced nephropathologist.

3 Results and Discussion

Style-Transfer Hyper Parameters. We elected to use the original VGG19 network during the style-transfer sample generation phase due to its light-weight yet powerful contextual representation ability [21]. Following Gatys *et al.*, we take the output of the fourth convolutional layer for content and the outputs of convolutional layers one through five for style [9]. We chose to initialize $\mathbf{x}_{out}$ to be an exact copy of $\mathbf{x}_{cont}$, allowing us to retain high content image fidelity while reducing the number of iterations needed to produce a meaningful output (100 iterations to generate each sample, as shown in Fig. 1). This allowed us

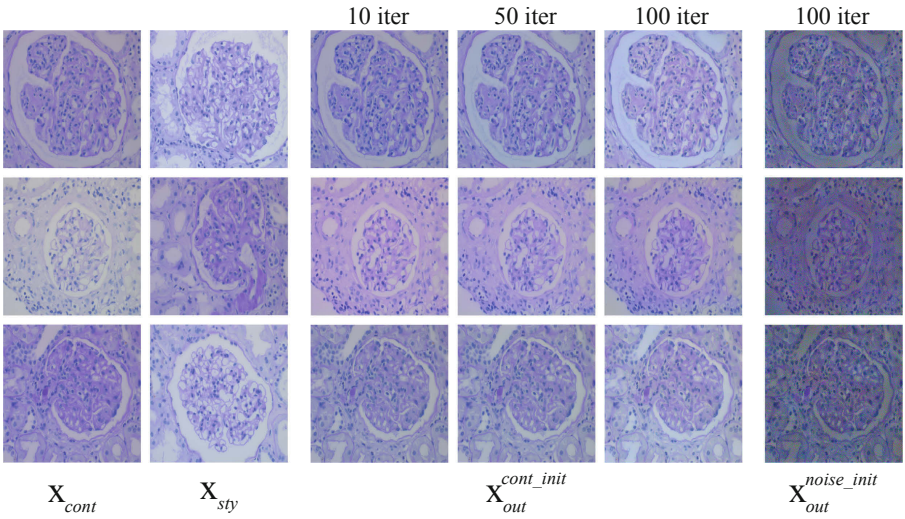

Fig. 1. Comparison of content and random initialization results. We note that output images initialized as the content image retain their morphological characteristics while capturing color and texture from the style image within 100 iterations. Output images initialized as noise appear distorted and discolored, failing to retain content fidelity.

to generate each image in 1.84 ± 0.03 s using a single GTX TITAN V gpu and pytorch, making it faster than other popular data augmentation techniques. This also helps soften the transfer of spatial relationships, allowing only slight structural modifications to occur that do not alter the desired target concept. We elected to use an α value of 2×10^{-4} following visual inspection of the generated samples by an experienced nephropathologist; this ensured that the generated samples retained the morphological characteristics of their respective content image while capturing the texture and color characteristics of their respective style image. When generating each $\mathbf{x}_{out}$, we selected one content and style image at random from the training set, irrespective of their associated label or section of origin. The generated style-transfer images were assigned the label of their corresponding content image and then appended to the training set before each augmented training experiment.

Bayesian Classification Analysis. During the classification task, we used the DenseNet-121 architecture pre-trained on the ImageNet dataset [12]. Both the original and StyPath augmented sample sets were resized to 256×256 pixels prior to being passed to each model; each classifier was trained for 200 epochs, using a batch size of 10, a drop rate of 0.1 in all bottleneck blocks, and a learning rate of 10^{-4}. Online augmentation was performed in all experiments for fair comparison between models; images had a 50% chance of being flipped horizontally, being flipped vertically, having up to 30% of their x and y axis cropped, and being rotated in either direction by up to 90°. Each testing fold in our five

fold cross validation scheme consisted of original glomerular images; no images derived from the same section could be present in both the training and testing set folds. We gradually added an equal amount of randomly selected AMR and non-AMR style-transfer images to each training set to identify where performance saturated. We note that the baseline model achieves a weighted Bayesian classification accuracy of 85.2% while the StyPath augmented model saturates around 88.5% (after adding 300 style-transfer samples to each class, as shown in Fig. 2a). To confirm that StyPath's performance increase had saturated, we also evaluated the model after adding 10000 style-transfer samples to each class, which yielded a Bayesian classification accuracy of 88.2%.

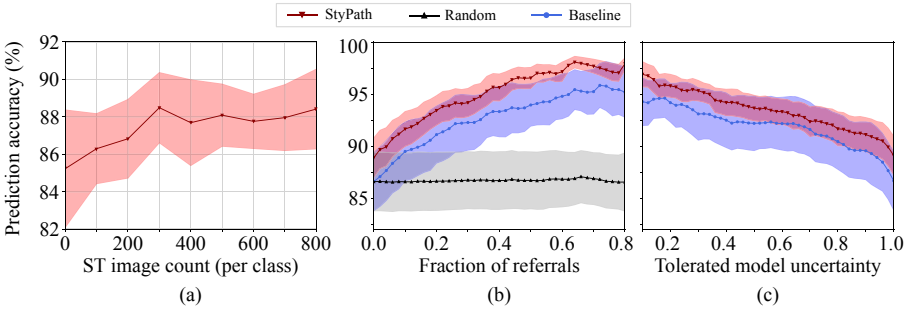

Fig. 2. Quantitative comparison of the baseline and StyPath augmented model performance. We note that the Bayesian performance of StyPath saturates after adding 300 Style-Transfer (ST) samples per class (a). When removing samples based on their uncertainty estimates, we note that the StyPath augmented model outperforms the baseline model while varying less across all five folds (b and c).

Epistemic Uncertainty Analysis. We then filtered samples based on their normalized epistemic uncertainty values; we observe that the StyPath augmented model consistently outperforms the baseline model while its performance varies significantly less across all folds (as shown in Fig. 2b and c). This result implies that StyPath augmented models have improved generalization ability, suggesting that the information being used by the model is more descriptive of the target concept.

Qualitative Analysis Using Grad-CAM. We note that certain StyPath augmented model predictions caused the Grad-CAM activations to shift, widen, or narrow down to diagnostic features present in the glomeruli, correcting erroneous baseline model predictions (as shown in Fig. 3, columns 1 through 5). While some failure cases indicate that the StyPath augmented model missed the desired concept (Fig. 3, columns 8 and 9), the initial label (non-AMR) of Fig. 3, column 6

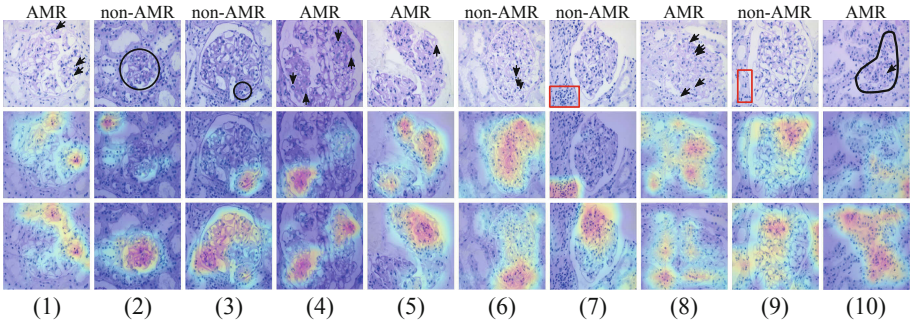

Fig. 3. Comparison of baseline (second row) and StyPath augmented (third row) Grad-CAM activations. Columns (1) through (5) represent success cases where the StyPath augmented prediction is correct and the baseline model prediction is incorrect, while columns (6) through (10) represent failure cases where the StyPath augmented prediction is incorrect and the baseline model prediction is correct. Black arrows in the original image (top row, labeled) are indicative of glomerulitis, mononuclear cells/infiltrates, and split glomerular basement membranes, while black and red outlines characterize regions of interest and misleading areas, respectively (Color figure online).

was put into question upon observing the StyPath augmented Grad-CAM activation; the model detects an accumulation of intracapillary mononuclear cells that are indicative of glomerulitis (and thus AMR). We observe that the baseline model correctly classifies Fig. 3, column 7 as non-AMR by counter intuitively focusing on an area of dense infiltrates in the periglomerular interstitium, which is not descriptive of non-AMR. Although the sample is misclassified, the StyPath augmented model instead focuses on the glomerular tuft, indicating that it detected the correct region of interest. We also note that generally, collapsed glomerular images seemed to pose a challenge for both the nephropathologist and the classifier, with both baseline and StyPath augmented models failing to focus on the areas indicative of AMR (Fig. 3, column 10).

4 Conclusions

We present a novel histological data augmentation technique called StyPath, which generates new histological samples through the sample and condition agnostic transfer of both spatial relationships and style. We use Bayesian inference to evaluate the technique and show that it improves both the performance and generalization ability of the classifier at a low computational cost. We then generated Grad-CAM representations of both the baseline and StyPath augmented models for assessment by an experienced nephropathologist. This assessment showed that the augmented model tended to focus on morphologically relevant information which ultimately improved its classification accuracy. Our future works aim to compare the performance of StyPath to other SOTA techniques on larger multi-conditional datasets.

Acknowledgments. This research was supported by the National Science Foundation (NSF-IIS 1910973).

References

1. Labelbox: labelbox (2020). https://labelbox.com
2. Adam, B., et al.: Banff initiative for quality assurance in transplantation (bifquit): reproducibility of polyomavirus immunohistochemistry in kidney allografts. Am. J. Transplant. **14**(9), 2137–2147 (2014)
3. Bejnordi, B.E., et al.: Stain specific standardization of whole-slide histopathological images. IEEE Trans. Med. Imaging **35**(2), 404–415 (2015)
4. BenTaieb, A., Hamarneh, G.: Adversarial stain transfer for histopathology image analysis. IEEE Trans. Med. Imaging **37**(3), 792–802 (2017)
5. Bug, D., et al.: Context-based normalization of histological stains using deep convolutional features. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 135–142. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_16
6. Dasari, S., Chakraborty, A., Truong, L., Mohan, C.: A systematic review of interpathologist agreement in histologic classification of lupus nephritis. Kidney Int. Rep. **4**(10), 1420–1425 (2019)
7. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158 (2015)
8. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)
9. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
10. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press, Cambridge (2016)
12. Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K.: Densenet: implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869 (2014)
13. Koelzer, V.H., et al.: Tumor budding in colorectal cancer revisited: results of a multicenter interobserver study. Virchows Arch. **466**(5), 485–493 (2015). https://doi.org/10.1007/s00428-015-1740-9
14. Macenko, M., et al.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110. IEEE (2009)
15. Martin, B., et al.: Interobserver variability in the h&e-based assessment of tumor budding in pt3/4 colon cancer: does it affect the prognostic relevance? Virchows Arch. **473**(2), 189–197 (2018)
16. Mobiny, A., Nguyen, H.V., Moulik, S., Garg, N., Wu, C.C.: Dropconnect is effective in modeling uncertainty of Bayesian deep networks. arXiv preprint arXiv:1906.04569 (2019)
17. Mobiny, A., Singh, A., Van Nguyen, H.: Risk-aware machine learning classifier for skin lesion diagnosis. J. Clin. Med. **8**(8), 1241 (2019)

18. Neal, R.M.: Bayesian Learning for Neural Networks. Lecture Notes in Statistics, vol. 118. Springer Science & Business Media, New York (2012). https://doi.org/10.1007/978-1-4612-0745-0
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
20. Shaban, M.T., Baur, C., Navab, N., Albarqouni, S.: Staingan: stain style transfer for digital histological images. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 953–956. IEEE (2019)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Tellez, D., et al.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. **58**, 101544 (2019)
23. Wilhelmus, S., et al.: Interobserver agreement on histopathological lesions in class III or IV lupus nephritis. Clin. J. Am. Soc. Nephrol. **10**(1), 47–53 (2015)