Proceedings of Machine Learning Research 116:10–24, 2020 Machine Learning for Health (ML4H) at NeurIPS 2019

Generative Image Translation for Data Augmentation in Colorectal Histopathology Images

Jerry Wei[1] JERRY.WENG.WEI@GMAIL.COM
Arief Suriawinata[2] ARIEF.A.SURIAWINATA@HITCHCOCK.ORG
Louis Vaickus[2] LOUIS.J.VAICKUS@HITCHCOCK.ORG
Bing Ren[2] BING.REN@HITCHCOCK.ORG
Xiaoying Liu[2] XIAOYING.LIU@HITCHCOCK.ORG
Jason Wei[1] JASON.20@DARTMOUTH.EDU
Saeed Hassanpour[1] SAEED.HASSANPOUR@DARTMOUTH.EDU
[1] *Dartmouth College, Hanover, NH, USA*
[2] *Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA*

Editors: Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

We present an image translation approach to generate augmented data for mitigating data imbalances in a dataset of histopathology images of colorectal polyps, adenomatous tumors that can lead to colorectal cancer if left untreated. By applying cycle-consistent generative adversarial networks (CycleGANs) to a source domain of normal colonic mucosa images, we generate synthetic colorectal polyp images that belong to diagnostically less common polyp classes. Generated images maintain the general structure of their source image but exhibit adenomatous features that can be enhanced with our proposed filtration module, called Path-Rank-Filter. We evaluate the quality of generated images through Turing tests with four gastrointestinal pathologists, finding that at least two of the four pathologists could not identify generated images at a statistically significant level. Finally, we demonstrate that using CycleGAN-generated images to augment training data improves the AUC of a convolutional neural network for detecting sessile serrated adenomas by over 10%, suggesting that our approach might warrant further research for other histopathology image classification tasks.

1. Introduction

Accurately analyzing medical images with deep learning classifiers often requires large, balanced datasets. For many diseases, however, the distribution of disease sub-classes in collected datasets is heavily skewed by each class's prevalence among patients, and so detecting rare diseases in medical images with deep learning can be challenging. In these situations, a reliable method of data augmentation can mitigate the effects of data imbalance by preventing overfitting and thus improving overall performance.

Previous work in data augmentation includes both traditional augmentation methods (rotations, flips, color jittering, etc.) and, more recently, generative models that synthesize completely new images. Since their development, generative adversarial networks (GANs) (Goodfellow et al., 2014), which use noise as an input variable, have been a popular method

of generating augmented data for improving image classification (Perez and Wang, 2017; Salehinejad et al., 2017). We hypothesized that, in the field of medical image analysis, data from one class might contain useful information to synthesize new data for another. As such, generative image translation models might suit this task better than models that do not account for information in other classes (e.g., models that use random noise as a basis for image generation).

In this paper, we present an image translation model for generating synthetic colorectal histopathology images. Since adenomatous preneoplastic polyps always originate from normal colonic mucosa, we use normal colonic mucosa as a source domain to generate synthetic images that are similar in structure but present adenomatous features. Our work makes the following contributions:

1. We demonstrate an image translation model that generates synthetic images of adenomatous colorectal polyps and propose a filtration module called Path-Rank-Filter that enhances the presence of adenomatous features in generated images.

2. We evaluate the quality of generated images through Turing tests with four gastrointestinal pathologists, finding that for the two adenomatous polyp classes tested, at least two of four pathologists could not distinguish between synthetic and real polyp images at a statistically significant level.

3. We show that using generated images as augmented data for training improves the AUC of a convolutional neural network in detecting sessile serrated adenomas by over 10%, indicating that our approach might be useful for other histopathology image classification tasks.

Our code for this project is publicly available.[1]

2. Related Work

Generative adversarial networks (GANs) have commonly been used in the field of medical image analysis. For magnetic resonance imaging (MRI) scans, Nie et al. (2016) used context-aware GANs to generate computed topography (CT) images from MRIs, and Yang et al. (2018) used conditional GANs (cGANs) to generate target modality MRIs given a particular source modality MRI. Furthermore, Dar et al. (2018) used cGANs to generate fake T1 and T2 MRIs and used an improved methodology by using end-to-end training of GANs that synthesize target images given source images. Hiasa et al. (2018) also translated MRIs to CT images with CycleGANs, adding a gradient consistency loss to encourage edge alignment between images. Salehinejad et al. (2018) used DCGANs to generate fake chest x-ray images from real ones, though the resulting fake images were at a lower resolution than real images, and Wang et al. (2018) used cGANs to reduce artifacts in CT images by learning to map an artifact-affected CT image to an artifact-free CT image.

In the field of histopathology in particular, many studies have used GANs for both image generation and image translation. Both Bayramoglu et al. (2018) and Rana et al. (2019) used cGANs to virtually stain Haemotoxylin and Eosin (H&E) lung tissue histopathology.

1. https://github.com/BMIRDS/HistoGAN

Similarly, Hou et al. (2017) and Quiros et al. (2019) generated fake histopathology samples with GANs, and Burlingame et al. (2018) used cGANs to translate pancreas tumors from H&E-stained to immunofluorescent. In terms of stain normalization, Bentaieb and Harmarneh (2017) used a GAN to normalize tissue samples in order to remove natural discolorations from tissue staining, and Cho et al. (2017) performed stain style transfer by replacing stain normalization models with cGANs. Moreover, Zanjani et al. (2018) integrated a Convolutional Neural Network (CNN) and Gaussian Mixture Model to jointly optimize the modeling and normalizing of color and intensity in H&E stained images.

In terms of data augmentation, both conventional methods and GANs have been used in previous research. Hussain et al. (2017) found that effective methods of data augmentation for images primarily include strategies such as flips, Gaussian noise, jittering, Gaussian blurring, and rotations, and Li et al. (2010) addressed class imbalances by oversampling abnormal classes and undersampling normal classes. For generative methods, Bass et al. (2019) synthesized augmented biomedical images with convolutional capsule GANs. Additionally, Gupta et al. (2019) used CycleGANs on x-ray images to generate augmented images of bone lesions, which were then added to a training set to improve a bone lesion classifier's AUC by 5%. Both papers, however, did not manually evaluate the quality of their generated images, and Gupta et al. (2019) did not have extensive ablation studies to provide insight on how their method could be applied to other datasets.

In our study, we apply CycleGAN to a colorectal histopathology image dataset to generate augmented data. We propose a filtration module called Path-Rank-Filter that improves the quality of generated images for some classes and perform extensive ablation studies. Furthermore, we evaluate our generated images manually with four pathologists and compare our CycleGAN model's ability to improve classifier performance with that of two other generative models: DCGAN (Radford et al., 2015) and DiscoGAN (Kim et al., 2017).

3. Image Translation in Colorectal Histopathology Images

Here, we discuss our approach for applying generative image translation to a dataset of colorectal histopathology images. We focus on cycle-consistent generative adversarial networks (Zhu et al., 2017) and propose a simple filtration module called Path-Rank-Filter that enhances the adenomatous features in generated images. Additionally, we describe the process of collecting our dataset as well as our experimental setup.

3.1. Cycle-Consistent Generative Adversarial Networks

We use a cycle-consistent generative adversarial network (CycleGAN) (Zhu et al., 2017) model to translate images of normal colonic mucosa to images of adenomatous colorectal polyps. Given two domains, X and Y, with training samples $\{x_i\}_{i=1}^{N}$, where $x_i \in X$, and $\{y_i\}_{i=1}^{N}$, where $y_i \in Y$, CycleGAN learns the mapping $G : X \rightarrow Y$ for unpaired image translation. For colorectal polyp images, we set X as normal colonic mucosa, which has many images, and Y as a less common polyp type with few images (e.g., tubular adenoma or sessile serrated adenoma) so that we can mitigate the imbalance of class Y by generating a set of augmented data $\{G(x_i)\}_{i=1}^{N}$ that presents features of domain Y.

3.2. Path-Rank-Filter

Because histopathology images differ in nature from images in standard computer vision datasets (e.g., MNIST or ImageNet), we propose a module called Path-Rank-Filter that improves CycleGAN's performance specifically for histopathology images. Whereas distinguishing between common classes in computer vision (e.g., cats and dogs) is relatively straightforward, histopathology images can contain a range of histologic features that determine whether an image can be classified as adenomatous. For instance, both an image with small amounts of tubular architectures and an image covered by tubular architectures would be classified by a pathologist as a tubular adenoma. We hypothesize that images with more prominent features will be more useful for training, and so instead of training a CycleGAN on the original $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$, we introduce the following filtration process (Figure 1):

1. We train a ResNet (He et al., 2015) f to classify X and Y. We define $f_Y(y_i)$ as the output probability of the ResNet for class Y when given image y_i as the input.
2. Then, we run the ResNet on all $\{y_i\}_{i=1}^N$. For some $\alpha \in (0, 1]$, we find $\{y\}_\alpha \subset \{y_i\}_{i=1}^N$ such that for all $y_i \in \{y\}_\alpha$, $f_Y(y_i)$ is in the highest α of all output probabilities $\{f_Y(y_i)\}_{i=1}^N$.
3. We train CycleGAN on $\{x_i\}_{i=1}^N$ and $\{y\}_\alpha$ instead of $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$.

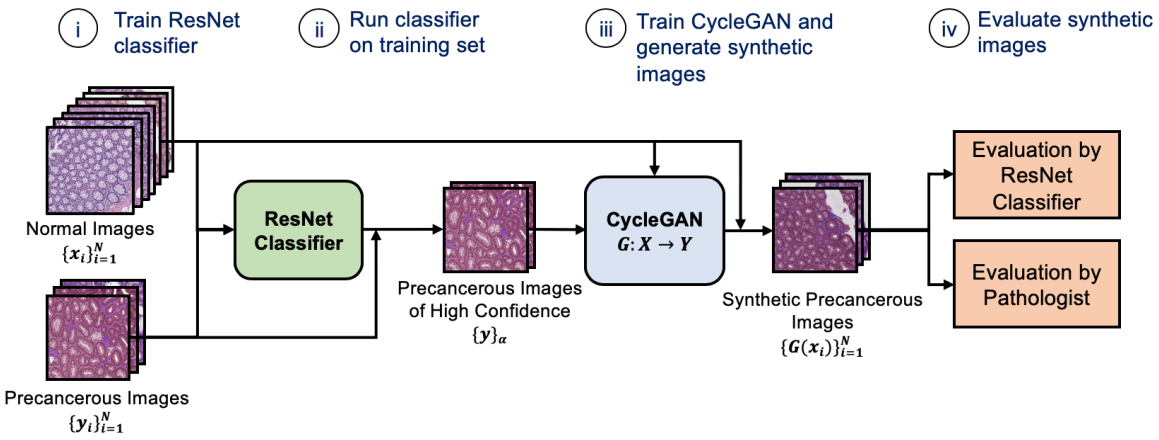


Figure 1: Process for generating synthetic histopathology images of rare colorectal polyp classes. Path-Rank-Filter (i-ii) enhances the adenomatous features in generated images by filtering the training data for CycleGAN for only images with strong adenomatous features.

Path-Rank-Filter uses the knowledge that an adenomatous class Y includes images with a range of histologic features. It thus finds the images with the strongest features that are most representative of class Y and uses those images to train CycleGAN.

3.3. Dataset Collection

Our dataset of colorectal polyp images was collected from the Dartmouth-Hitchcock Medical Center in New Hampshire, USA, our tertiary medical institution. We collected 427 high-resolution whole-slide images, which we split into a training set of 326 whole-slide images and a testing set of 101 whole-slide images. For the training set, pathologists annotated all whole-slide images with bounding boxes representing regions of interest, for a total of 3517 variable-size image crops. Each image crop was labeled with a single class for the polyp type, which was either benign (normal or hyperplastic), or adenomatous (tubular adenoma, tubulovillous/villous adenoma, or sessile-serrated adenoma). The distributions of different classes in our training set is shown in Figure 2.

For the testing set, pathologists annotated the whole-slide images for fixed-size tiles of classic examples of polyp types (224×224 pixels), and polyp type labels were verified by two pathologists so that our evaluation was as close to ground truth as possible. Our final testing set, which is used in section 4.3, had 261 hyperplastic polyp images and 39 sessile serrated adenoma images.

3.4. Experimental Setup and Motivation

In this study, we set tubular adenoma (TA) and sessile serrated adenoma (SSA), two adenomatous polyp types that respectively account for only 14.8% and 3.3% of our dataset by size, as the target domains for data generation. As a source domain, we use normal colonic mucosa images, since both tubular and sessile serrated adenomas emerge as a result of cytological transformations on normal colonic mucosa. For all classifiers, we use the ResNet architecture (He et al., 2015) and train each classifier for 20 epochs. We conducted an ablation study for our particular classification task and found that increasing the depth of the neural network did not substantially improve performance (Supplementary Figure 1). Thus, for all experiments, we used the model with the lowest number of parameters, ResNet-18, so that experiments can be replicated more quickly.

4. Experiments

We perform extensive experiments to evaluate the ability and usefulness of generative image translation on colorectal polyp histopathology images. We measure the strength of our filtering method using a pre-trained classifier, finding that CycleGAN with Path-Rank-Filter generates images that are substantially closer to the target domain (i.e. exhibit more adenomatous features) than when Path-Rank-Filter is not used. Next, we perform a clinical evaluation of our images by conducting a Turing test with four gastrointestinal pathologists, finding that three of the four pathologists could not differentiate at least half of the synthetic images from real images. Finally, we evaluate how adding the generated images as augmented data for training a ResNet classifier can improve performance for detecting sessile serrated adenomas, a clinically important distinction in colorectal cancer screening.

While we limit the scope of this paper to a single source domain, normal colonic mucosa, we show qualitative results of experiments on other source domains in Supplementary Figure 6.

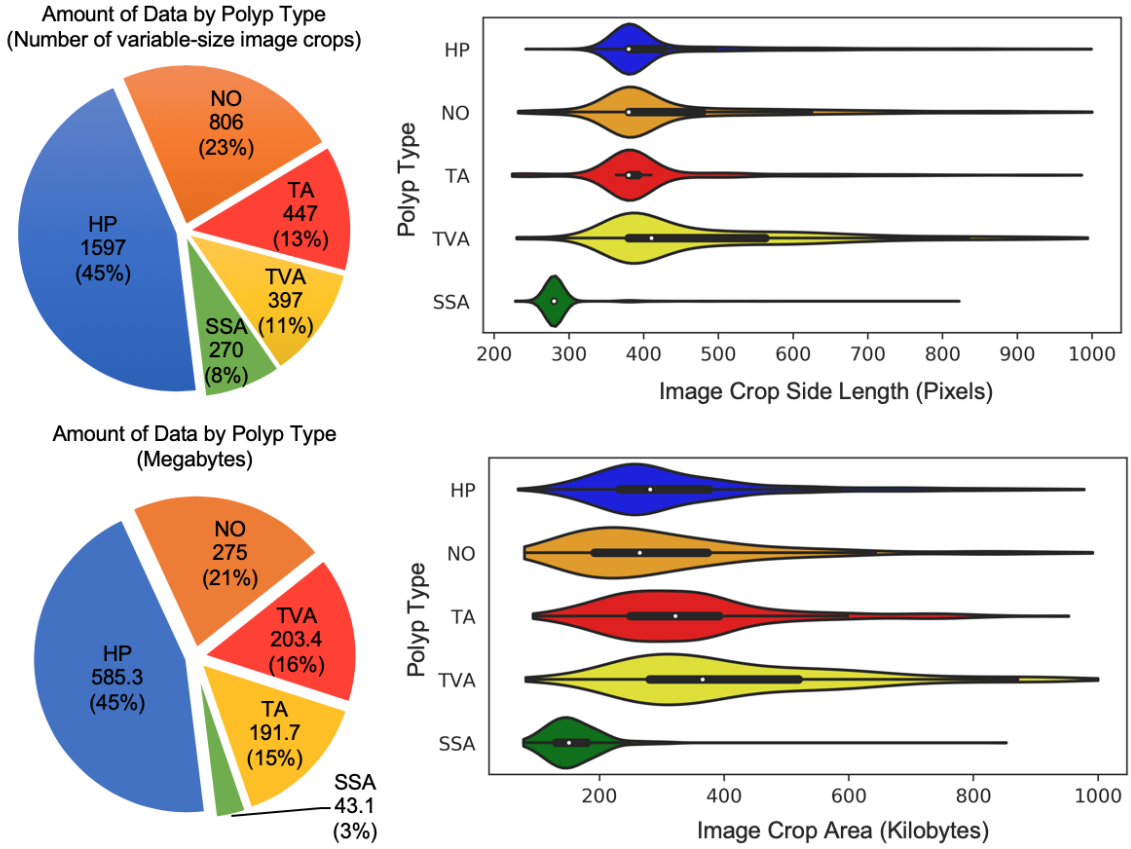


Figure 2: Distribution for collected dataset of colorectal polyp histopathology images. HP: hyperplastic polyp, NO: normal colonic mucosa, TVA: tubulovillous/villous adenoma, TA: tubular adenoma, SSA: sessile serrated adenoma. Two diagnostically relevant classes of adenomatous polyps, tubular adenoma (TA) and sessile serrated adenoma (SSA), comprise only 14.8% and 3.3% of the dataset, respectively.

4.1. Enhancing Adenomatous Features with Path-Rank-Filter

In this experiment, we evaluate how Path-Rank-Filter can select a subset of the adenomatous training images with the strongest adenomatous features for CycleGAN so that the generated images will also have a strong presence of features representing the desired class. For the three adenomatous classes of polyps (tubular, tubulovillous/villous, and sessile serrated), we apply CycleGAN using Path-Rank-Filter with filtration parameter values of $\alpha = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ on the 9054 normal colonic mucosa images in our training set to generate 9054 images of the target adenomatous class. We then measure the prominence of adenomatous features in our generated images by using a pre-trained classifier to evaluate the percent of generated images that were actually classified as the intended target class (Table 1).

Polyp Class	$\alpha = 1$	$\alpha = 1/2$	$\alpha = 1/4$	$\alpha = 1/8$	$\alpha = 1/16$	$\alpha = 1/32$
TA	35.4	64.4	79.6	87.6	89.2	**93.8**
TVA	32.7	67.3	49.4	63.1	85.9	**86.1**
SSA	37.0	20.9	21.5	38.5	23.4	43.7

Table 1: Percent of synthetic images generated by a CycleGAN with various α parameters for Path-Rank-Filter that were classified by a pre-trained classifier as the intended class. 9054 synthetic images were evaluated for each class and α value. TA: tubular adenoma, TVA: tubulovillous/villous adenoma, SSA: sessile serrated adenoma.

Based on this evaluation metric, Path-Rank-Filter substantially enhanced adenomatous features in generated images for TA and TVA. For these two classes, the highest classification performance was at $\alpha = \frac{1}{32}$, with the pre-trained classifier correctly detecting 93.8% of generated images for TA and 86.1% for TVA. These high accuracies seem to reflect the nature of these two adenomatous classes, for which images in the training set reflect a range of features. TA images are defined by hyperchromatic, pencillate nuclei; pathologists will label both images with small hints of pencillate nuclei and obviously strong tubular features as tubular. Of the same nature, TVA images are characterized by finger-like extensions with hyperchromatic, pencillate nuclei, and therefore some images will have more villous features than others.

For SSA, on the other hand, Path-Rank-Filter did not significantly improve the performance. We hypothesize that this result reflects the differing nature of SSAs, which are classified by the presence of a single broad-based crypt. Unlike TAs and TVAs, SSAs do not present a spectrum of histological features, and so it makes sense that Path-Rank-Filter does not choose a better subset of SSAs for training CycleGAN, and therefore generated images did not exhibit stronger features of SSAs.

Furthermore, we select example images to examine the histologic features as we use different filtration parameters (Figure 3). For TA, we see that CycleGAN transforms normal crypts by introducing pencillate nuclei into the crypt borders, altering cell color, and merging small crypts into more complex structures. For TVA, crypts become more elongated and finger-like for smaller α parameters. For SSA, however, the quality of adenomatous features did not substantially improve with smaller α parameters; perhaps the SSA example shown when using all images for training $\alpha = 1$ has the strongest features, although interpretations might differ among pathologists. More examples of generated images for varying α are shown in Supplementary Figures 2 (TA), 3 (TVA), and 4 (SSA). Generated images of tubular adenomas after various epochs are shown in Supplementary Figure 5.

4.2. Evaluation by Pathologists

We further measure the quality of generated adenomatous images through clinical evaluation by four gastrointestinal pathologists. For the tubular and sessile serrated classes,[2] the two least common classes in our dataset, we presented the four pathologists with a set of

2. Manual evaluation is costly, and so we do not evaluate tubulovillous/villous adenoma in this paper.

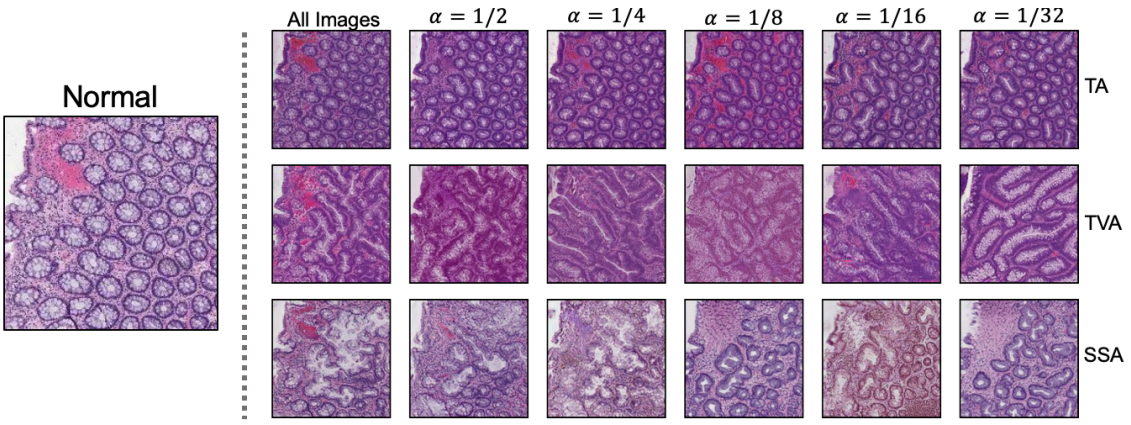


Figure 3: CycleGAN's generated images for different values of α. For instance, $\alpha = 1/4$ means that the top 25% of images with the highest output probabilities from a ResNet were used to train CycleGAN. TA: tubular adenoma, TVA: tubulovillous/villous adenoma, SSA: sessile serrated adenoma. For TA and TVA, adenomatous features were enhanced at smaller α values.

200 unlabeled images: 100 real images and 100 generated (fake) images. Each pathologist independently classified each image as either real or fake. As shown in Figure 4, at least half of the pathologists could not distinguish real and fake images at a statistically significant level, correctly distinguishing some fake images while also incorrectly labeling real images as fake.

We also perform statistical analysis on the pathologists' overall accuracies, using $x_0 = 0.5$ as the expected accuracy for random guessing and each pathologist's accuracy on the $n = 200$ images as $\hat{x}$ to calculate the z-score for each pathologist (Equation 1).

$$z = \frac{\hat{x} - x_0}{\sqrt{\frac{x_0(1-x_0)}{n}}} \tag{1}$$

We then calculate p for each pathologist given the null hypothesis $H_0 : \hat{x} = x_0$. With this configuration, a p-value where $p < 0.05$ is statistically significant (i.e., the pathologist is able to distinguish between real and fake images).

For tubular adenoma images, only one pathologist was able to differentiate real images from synthetic images at a statistically significant level. For sessile serrated adenoma images, two pathologists were able to distinguish between real and synthetic images at a statistically significant level. Based on feedback from pathologists, fake sessile serrated adenoma images were easier to identify because our CycleGAN model created a subtle mosaic-like pattern in the whitespace of images. Sessile serrated adenomas tended to have more whitespace because they are defined by a single large crypt (of mostly whitespace), which might explain why it was easier to detect fake sessile serrated adenomas than tubular adenomas.

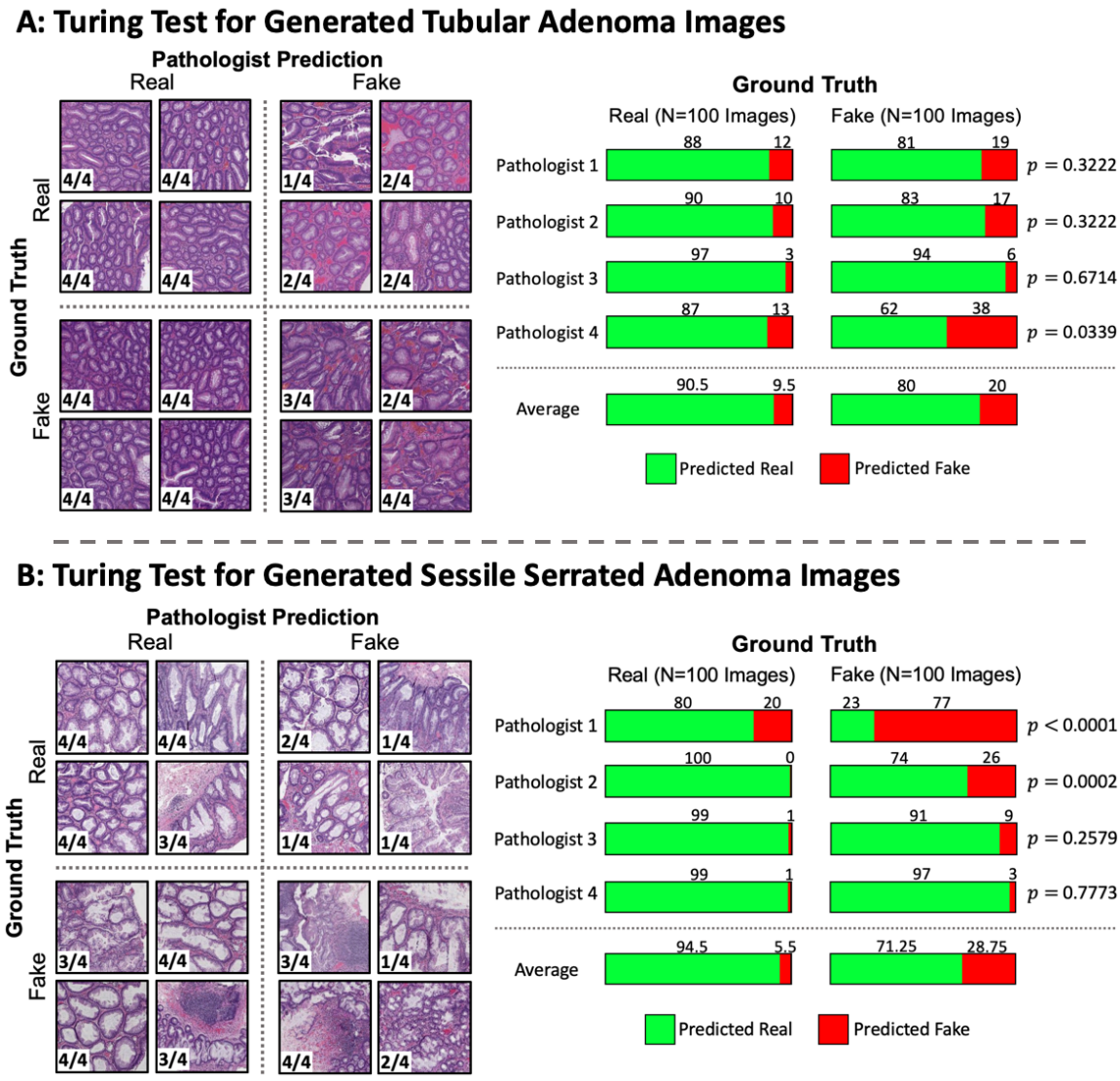


Figure 4: Results of Turing test for whether gastrointestinal pathologists could distinguish real and fake images of tubular adenomas (A) and sessile serrated adenomas (B). Left: example real and generated images that were classified correctly and incorrectly by pathologists, with the number of pathologists who labeled the image as such denoted in the lower left corner. Right: evaluation of real and fake images by four pathologists.

4.3. Improving Classifier Performance

Image translation can mitigate class imbalances in training sets by generating synthetic images of rare classes. We generated synthetic images of sessile serrated adenomas (only represented by 3% of the training set) and used them as augmented data for training a

ResNet classifier to distinguish between hyperplastic polyps (benign) and sessile serrated adenomas (adenomatous), a clinically important task in colorectal cancer screening (Korbar et al., 2017a,b). We applied CycleGAN to all 9054 normal colonic mucosa images in our training set to generate 9054 images of the sessile serrated class, and added these images into the training set. Then, we used this dataset for training a ResNet and evaluated it on a test set of 261 hyperplastic polyp images and 39 sessile serrated adenoma images, comparing our ResNet's performance with that of ResNets trained on generated data from DiscoGAN and DCGAN, as well as ResNets trained without augmented data (Figure 5A). Including CycleGAN-generated images for training boosted classification AUC by over 10%, outperforming DCGAN-generated images and DiscoGAN-generated images.

We also train ResNet on a training set consisting of the same real hyperplastic images but with synthetic images as the only available sessile serrated adenoma images (Figure 5B). Once again, the model trained on CycleGAN-generated images outperformed the models trained on DCGAN-generated images and DiscoGAN-generated images by 8% and 23%, respectively. In both experiments, the ResNet that was trained using CycleGAN-generated images achieved the highest AUC.

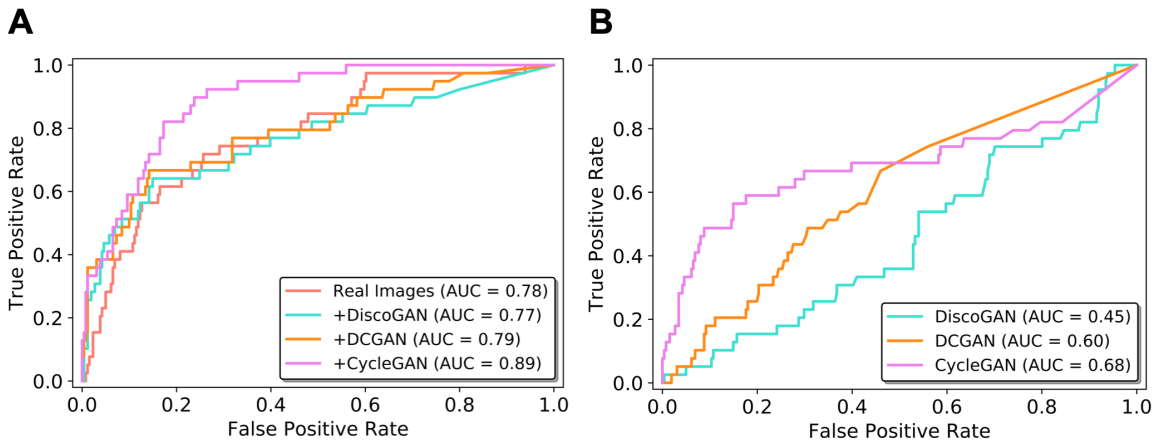


Figure 5: **A**: AUCs of ResNets trained on real images with synthetic images from different generative models given as additional training data. **B**: AUCs of ResNets trained without real images and with synthetic images from different generative models as the only available training data. In both experiments, the ResNet that was trained with CycleGAN's synthetic images had the highest AUC.

5. Limitations and Discussion

Although we show some promising results in terms of image quality and ability to improve the performance of a ResNet classifier, our study has notable limitations. First, fair manual evaluation of images is non-trivial. Even though the pathologists in our study have years of experience examining colorectal polyp slides, these Turing tests do not perfectly reflect image quality, since pathologists do not distinguish real and fake data as a task in clinical

practice. Furthermore, variation in results suggest that distinguishing fake images might depend highly on the individual pathologists, and some pathologists reported that they could better distinguish real and fake images as they saw more images. Finally, we only showed pathologists fixed-sized tiles of images; generating an entire high-resolution slide with high-quality features is a substantially more challenging task.

In terms of improving classifier training, we had hoped that training with synthetic data would achieve the same performance as training with real data, but a ResNet trained on only synthetic SSA images achieved an AUC of only 0.68 (Figure 5), much lower than the AUC of a classifier trained on both real and synthetic data (0.89). This result suggests that although the quality of a single generated image might be comparable to that of a single real image, the quality of the set of generative images likely does not match that of a set of real images.

Our paper has explored image translation for data augmentation in colorectal histopathology images. Whereas most work in generative data augmentation focuses on generating images from random noise, we note that images from other classes might be helpful in the field of histopathology and therefore take an image translation approach. Future work might include evaluating our method on other datasets to evaluate the generalizability of our approach.

Acknowledgments

This research was supported in part by National Institute of Health grants R01LM012837 and P20GM104416. We thank Naofumi Tomita for helpful feedback on the study design.

References

Cher Bass, Tianhong Dai, Benjamin Billot, Kai Arulkumaran, Antonia Creswell, Claudia Clopath, Vincenzo De Paulo, and Anil Anthony Bharath. Image synthesis with a convolutional capsule generative adversarial network. In *Proceedings of the International Conference on Medical Imaging with Deep Learning*, 2019. https://openreview.net/pdf?id=rJen0zC1lE.

Neslihan Bayramoglu, Mika Kaakinen, Lauri Eklund, and Janne Heikkila. Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 64–71, 2018. https://ieeexplore.ieee.org/document/8265226.

Aicha Bentaieb and Ghassan Harmarneh. Adversarial stain transfer for histopathology image analysis. In *Proceedings of IEEE Transactions on Medical Imaging*, volume 37, 2017. https://ieeexplore.ieee.org/document/8170242.

Erik A. Burlingame, Adam Margolin, Joe Gray, and Young Hwan Chang. SHIFT: Speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks. In *Proceedings of IEEE Transactions on Medical Imaging*, volume 10581, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6166432/.

Hyungjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. Neural stain-style transfer learning using GAN for histopathlogical images. In *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 2017. https://arxiv.org/pdf/1710.08543.pdf.

Salman UH. Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cuker. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE Transactions on Medical Imaging*, 2018. https://arxiv.org/pdf/1802.01221.pdf.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2, pages 2672–2680, 2014. https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Anant Gupta, Srivas Venkatesh, Sumit Chopra, and Christian Ledig. Generative image translation for data augmentation of bone legion pathology. *arXiv*, 2019. https://arxiv.org/pdf/1902.02248.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385. https://dblp.org/rec/bib/journals/corr/HeZRS15.

Yuta Hiasa, Yoshito Otake, Masaki Takao, Takumi Matsuoka, Kazuma Takashima, Aaron Carass, Jerry L. Prince, Nobuhiko Sugano, and Yoshinobu Sato. Cross-modality image synthesis from unpaired data using CycleGAN. *arXiv*, 2018. https://arxiv.org/pdf/1803.06629.pdf.

Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M. Kurc, Rajarsi R. Gupta, and Joel H. Saltz. Unsupervised histopathology image synthesis. *arXiv*, 2017. https://arxiv.org/pdf/1712.05021.pdf.

Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA Annual Symposium Proceedings*, 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977656/.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv*, 2017. https://arxiv.org/pdf/1703.05192.pdf.

Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Caherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, and Saeed Hassanpour. Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In *2017 IEEE Confrence on Computer Vision and Pattern Recognition Workshops*, 2017a. https://ieeexplore.ieee.org/document/8014848.

Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Catherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, and Saeed Hassanpour. Deep learning for classification of colorectal polyps on whole-slide images. *Journal of Pathology Informatics*, 8(1), 2017b. http://www.jpathinformatics.org/article.asp?issn=2153-3539;year=2017;volume=8;issue=1;spage=30;epage=30;aulast=Korbar.

Der-Chiang Li, Chiao-Wen Liu, and Susan C. Hu. A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine*, 40:509–518, 2010. https://www.ncbi.nlm.nih.gov/pubmed/20347072.

Dong Nie, Roger Trullo, Caroline Petitjean, Su Ruan, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. *arXiv*, 2016. https://arxiv.org/pdf/1612.05362.pdf.

Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621, 2017. URL http://arxiv.org/abs/1712.04621. https://dblp.org/rec/bib/journals/corr/abs-1712-04621.

Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. PathologyGAN: Learning deep representations of cancer tissue. *arXiv*, 2019. https://arxiv.org/pdf/1907.02644.pdf.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*, 2015. https://arxiv.org/pdf/1511.06434.

Aman Rana, Gregory Yaunery, Alarice Lowe, and Pratik Shah. Computational histological staining and destaining of prostate core biopsy RGB images with generative adversarial neural networks. *arXiv*, 2019. https://arxiv.org/pdf/1811.02642.pdf.

Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. *CoRR*, abs/1712.01636, 2017. URL http://arxiv.org/abs/1712.01636. https://dblp.org/rec/bib/journals/corr/abs-1712-01636.

Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. Generalization of deep neural networks for chest pathology classification in X-Rays using generative adversarial networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018. https://arxiv.org/pdf/1712.01636.pdf.

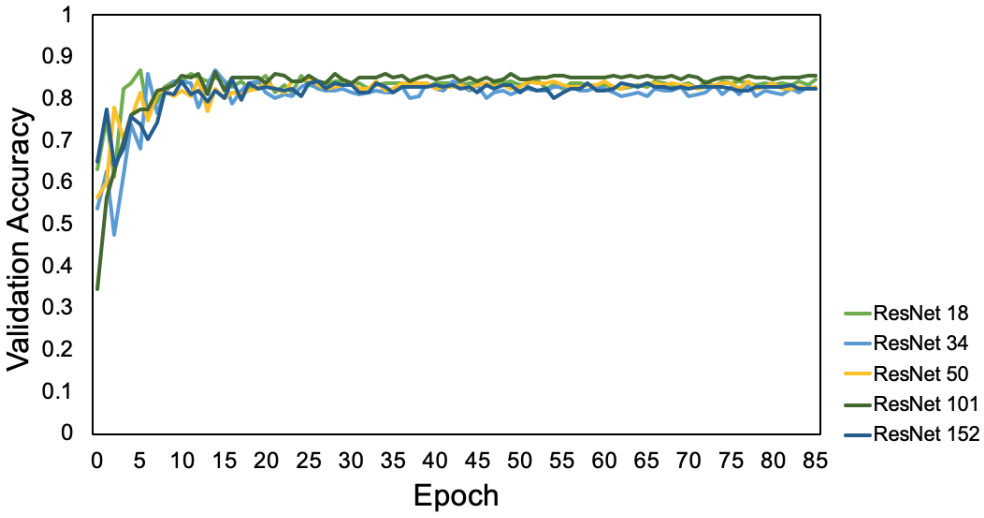
Jianing Wang, Yiyuan Zhao, Jack H. Noble, and Benoit M. Dawant. Condtional generative adversarial networks for metal artifact reduction in CT images of the ear. In *Proceedings of Medical Imaging Computing and Computer Assisted Inventions*, pages 3–11, 2018. https://link.springer.com/chapter/10.1007/978-3-030-00928-1_1.

Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I-Chao Chang, and Yan Xu. MRI image-to-image translation for cross-modality image registration and segmentation. *arXiv*, 2018. https://www.semanticscholar.org/paper/MRI-Image-to-Image-Translation-for-Cross-Modality-Yang-Li/572043ee8a930ab9ef76bca236b5593f83877985.

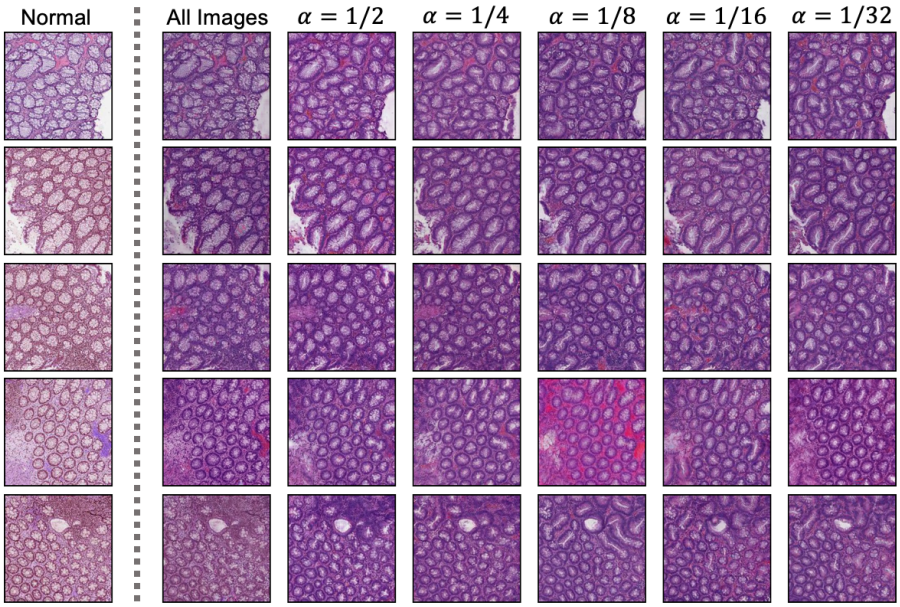
Farhad Ghazvinian Zanjani, Svitlana Zinger, and Peter H. N. de With. Deep convolutional gaussian mixture model for stain-color normalization of histopathological images. In *Proceedings of Medical Imaging Computing and Computer Assisted Inventions*, pages 274–282, 2018. https://link.springer.com/chapter/10.1007/978-3-030-00934-2_31.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. http://arxiv.org/abs/1703.10593.

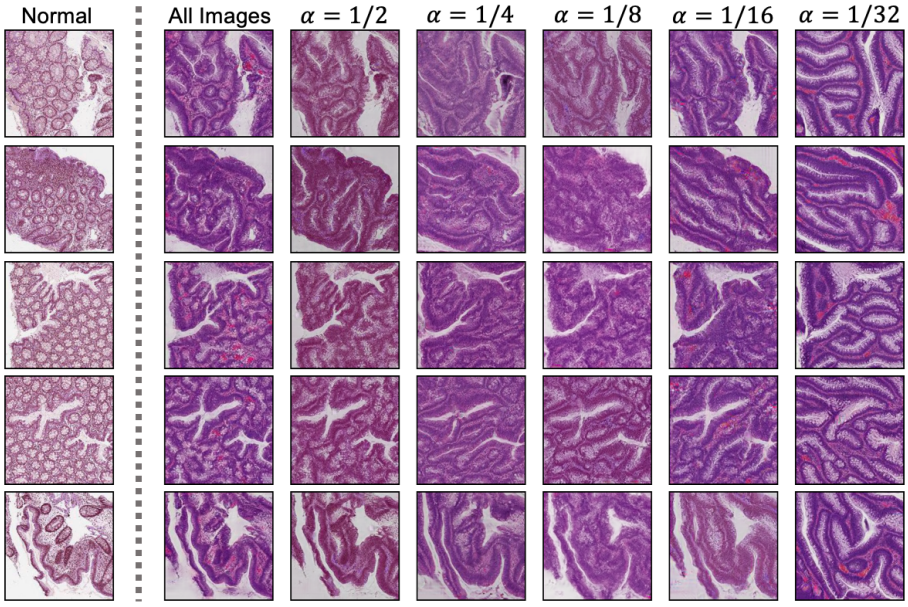
Supplementary Figures



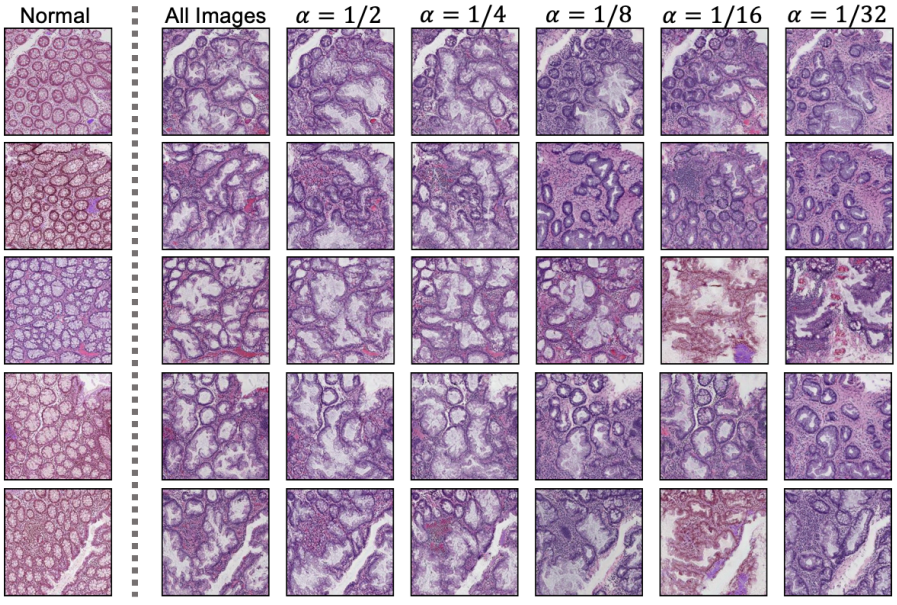
Supplementary Figure 1: Validation accuracy of ResNet classifiers of varying depth. Performance did not improve substantially for deeper networks.



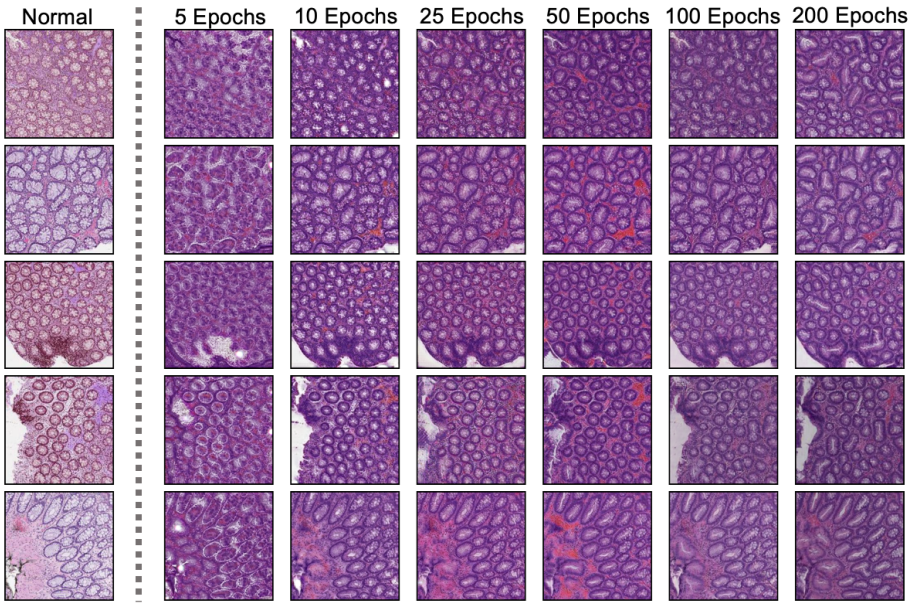
Supplementary Figure 2: Examples of tubular adenoma images generated by CycleGAN with Path-Rank-Filter at varying α levels. Adenomatous features were enhanced at lower α.



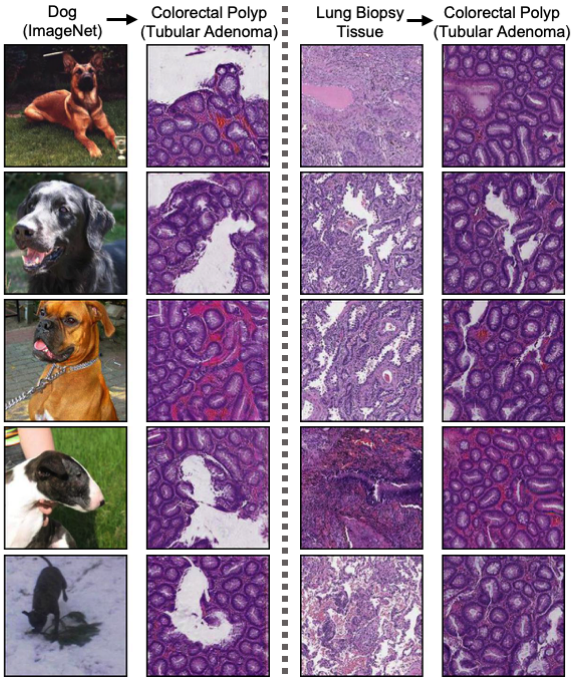
Supplementary Figure 3: Examples of tubullovillous/villous adenoma images generated by CycleGAN with Path-Rank-Filter at varying α levels. Adenomatous features were enhanced at lower α.



Supplementary Figure 4: Examples of sessile serrated adenoma images generated by CycleGAN with Path-Rank-Filter at varying α levels. Using lower α values did not enhance the features of seesile serrated adenomas.



Supplementary Figure 5: Examples of tubular adenoma images generated by CycleGANs trained for 5, 10, 25, 50, 100, and 200 epochs. Convergence occurred at approximately 200 epochs.



Supplementary Figure 6: Tubular adenoma images generated with CycleGAN using dogs from ImageNet and lung biopsy tissue samples as source domains.