arXiv:2102.01678v2 [eess.IV] 3 Jun 2021

Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation

Rikiya Yamashita, Jin Long, Snikitha Banda, Jeanne Shen, Daniel L. Rubin
Stanford University
{rikiya, jinlong, jeannes, rubin}@stanford.edu

ABSTRACT

Suboptimal generalization of machine learning models on unseen data is a key challenge which hampers the clinical applicability of such models to medical imaging. Although various methods such as domain adaptation and domain generalization have evolved to combat this challenge, learning robust and generalizable representations is core to medical image understanding, and continues to be a problem. Here, we propose **STRAP** (**S**tyle **TR**ansfer **A**ugmentation for histo**P**athology), a form of data augmentation based on random style transfer from non-medical style source such as artistic paintings, for learning domain-agnostic visual representations in computational pathology. Style transfer replaces the low-level texture content of an image with the uninformative style of randomly selected style source image, while preserving the original high-level semantic content. This improves robustness to domain shift and can be used as a simple yet powerful tool for learning domain-agnostic representations. We demonstrate that STRAP leads to state-of-the-art performance, particularly in the presence of domain shifts, on two particular classification tasks in computational pathology.

1 Introduction

While deep learning has demonstrated remarkable performance on medical imaging tasks over the past few years, the performance drop usually observed when generalizing from internal to external test data remains a key challenge in the medical application of machine learning models. Supervised learning assumes that training and testing data are sampled from the same distribution, *i.e.*, in-distribution, whereas in practice, the training and testing data typically originate from related domains, but which follow different distributions, *i.e.*, out-of-distribution. This phenomenon, known as domain shift [1], hampers the clinical applicability of such models, especially when the annotated datasets are limited in size or the target domain is highly heterogeneous.

One approach to tackling this domain shift problem is domain adaptation, which learns to align the feature distribution of the source domain with that of the target domain in a domain-invariant feature space. However, domain adaptation typically requires access to at least a few data samples from the target domain during training, which is not always available for medical applications. Another approach is domain generalization, which aims to adapt from multiple labeled source domains to an unseen target domain without needing to access data samples from the target domain. However, domain generalization typically requires multi-source training setting. Additionally, these approaches assume the target data are homogeneously sampled from the same distribution, an unrealistic scenario in most real-world medical applications, where models must deal with mixed-domain data (*e.g.*, scanner, protocols, medical sites) without their domain labels. In the present study, our focus is to address a challenging yet practical problem of knowledge transfer from one labeled source domain to multiple target domains, a task referred to as domain agnostic learning [2] or single-domain generalization [3], where we train the model on source data from a single domain and generalize it to

Correspondence to rikiya@stanford.edu
Code: https://github.com/rikiyay/style-transfer-for-digital-pathology

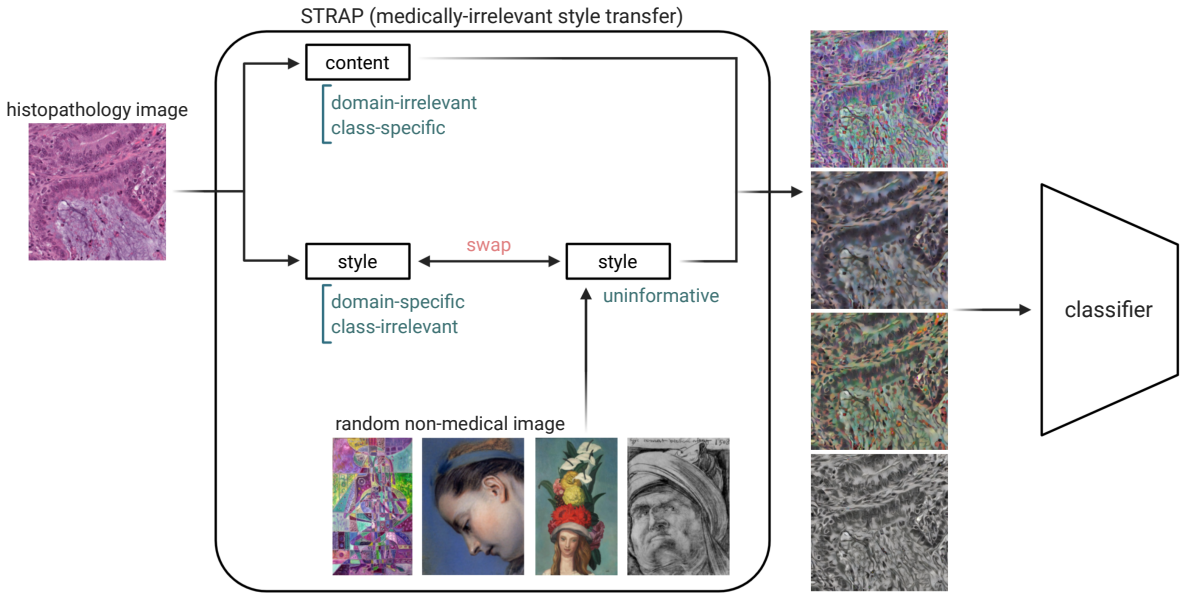

Figure 1: Overview of STRAP

unseen target data from multiple domains. A solution to domain-agnostic learning/single-domain generalization should learn domain-invariant and class-specific visual representations, as humans do.

Geirhos *et al*. [4] showed that 1) convolutional neural networks (CNNs) trained on the ImageNet dataset are biased towards texture, whereas humans are more reliant on global shape for distinguishing classes, 2) CNNs tend not to cope well with domain shifts, *i.e*., the change in image statistics from those on which the networks have been trained to those which the networks have never seen before, and 3) increasing shape bias by training on a stylized version of the ImageNet generated using style transfer improves accuracy, robustness, and generalizability.

Neural style transfer [5] refers to a CNN-based image transformation algorithm that manipulates the low-level texture representation of an image, *i.e*., style, while preserving its semantic content. The original method by Gatys *et al*. uses Gram matrices of the activations from different layers of a CNN to represent the style of an image. Then it uses an iterative optimization method to generate a new image from white noise by matching the activations with the content image and the Gram matrices with the style image. Huang and Belongie later proposed an improved approach called adaptive instance normalization (AdaIN) [6], which aligns the mean and variance of the content features with those of the style features. AdaIN enables arbitrary style transfer in real-time. Jackson *et al*. [7] demonstrated that, in computer vision tasks for natural images, data augmentation via style transfer with randomly selected artistic paintings as a style source improves robustness to domain shift, and can be used as a simple, domain-agnostic alternative to domain adaptation.

In medical imaging, machine learning models often suffer from domain shift in test data caused by heterogeneity from various sources, such as scanners, protocols, and medical sites. We know that human experts, such as radiologists and pathologists, are able to learn domain-agnostic visual representations and, thus, generalize across domains, particularly in the presence of domain shifts. We postulate that 1) human experts in medical imaging are also biased towards shape rather than texture as Geirhos *et al*. demonstrated [4], and 2) the low-level texture content of an image tends to be domain-specific, leading to suboptimal performance of deep learning models on domain-shifted unseen data, whereas high-level semantic content is more domain-invariant, from which ubiquitous class-specific visual representations can be learned.

Here, we propose **STRAP** (**S**tyle **TR**ansfer **A**ugmentation for histo**P**athology), a form of data augmentation based on random style transfer with non-medical style source, as a solution to learning domain-agnostic visual representation, particularly in computational pathology (Figs. 1, 2). In this study, the term "domain" refers to scanners, stain and scan protocols, and, more broadly, medical sites. We introduce STRAP as a solution to domain agnostic learning (*i.e*., single-domain generalization), and then, further assess its efficacy on conventional domain generalization setting

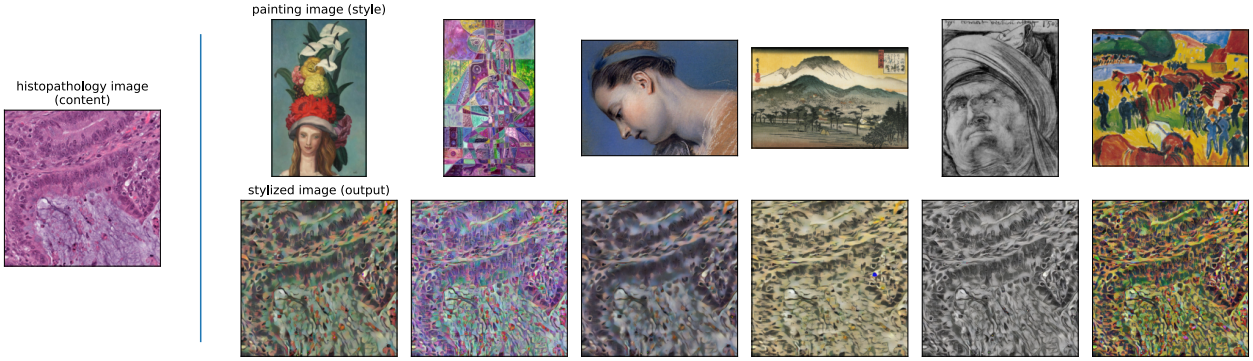

Figure 2: Style transfer with artistic paintings as a style source (stylization coefficient of 1.0) applied to a histopathology image (content on the left). Overall geometry is preserved, but the style, including texture, color, and contrast, is replaced with an uninformative style of a randomly selected artistic painting.

(*i.e.*, multi-source domain generalization). More specifically, we studied the proposed approach on two classification tasks in different domain generalization scenarios. The first is classifying colorectal cancer into two distinct genetic subtypes based on microsatellite status using hematoxylin and eosin (H&E)-stained, formalin-fixed, paraffin embedded (FFPE) whole-slide images (WSIs) of surgically resected colorectal cancers in single-domain generalization setting (models are trained on a single-domain dataset and tested on a mixed-domain dataset), hereafter referred to as genetic subtype classification task. The second is classifying presence or absence of breast cancer metastases in image patches extracted from histopathlogic scans of lymph node sections in multi-source domain generalization setting (models are trained on a multi-source domain dataset and tested on a single-domain dataset), hereafter referred to as tumor identification task. We compare STRAP against two standard baseline methods widely used in computational pathology, stain normalization [8] and stain augmentation [9], both of which apply medically-relevant transformation to the source images, whereas STRAP performs medically-irrelevant transformation.

We studied the effect of difference in style source by using artistic paintings, natural imaging, and histopathologic imaging as style sources (the former two apply medically-irrelevant style transfer, whereas the latter applies medically-relevant style transfer), and the effect of difference in stylization coefficient on the STRAP performance. Moreover, to gain insights into the differences in learning dynamics among the three approaches (STRAP, stain normalization, and stain augmentation), we performed following three experiments on the genetic subtype classification task: 1) we tested model performance on stylized version of the out-of-distribution test data; 2) we evaluated differential responses to the low-frequency components of the out-of-distribution test data; and 3) we visualized saliency maps on the low-frequency components of the out-of-distribution test data using integrated gradients [10]. The latter two experiments were inspired by Wang *et al*. [11], who showed that 1) CNNs can exploit high-frequency image components which humans do not consciously perceive and 2) models which exploit low-frequency components generalize better than those which exploit the high-frequency spectrum.

Our contributions are summarized as follows: 1) we present STRAP, a form of medically-irrelevant data augmentation based on random style transfer for computational pathology; 2) we utilize STRAP to improve both single-domain and multi-source domain generalization for two classification tasks in computational pathology; and 3) our experiments suggest that STRAP helps models learn from semantic contents and low-frequency components of the data, on which humans tend to rely in recognizing objects [12].

2 Methods

2.1 Style transfer augmentation with non-medical style source (STRAP)

Inspired by Geirhos *et al*. [4] and Jackson *et al*. [7], we propose STRAP, a form of medically-irrelevant data augmentation based on random style transfer for computational pathology, which replaces the style of the histopathology image (including texture, color, and contrast) with an uninformative style of a randomly selected non-medical image, while predominantly preserving the semantic content (global object shapes) of the image. We hypothesize that the style of the histopathology images is domain-specific and class-irrelevant, whereas the semantic content is domain-irrelevant and class-specific; therefore, STRAP facilitates learning domain-agnostic representations.

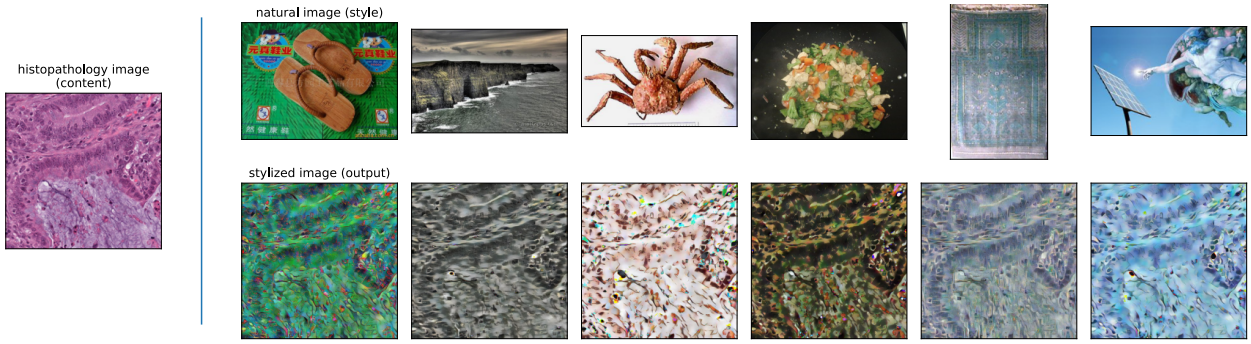

Figure 3: Style transfer with the Natural Imaging style source applied to a histopathology image (content on the left). Overall geometry is preserved, but the style, including texture, color, and contrast, is replaced with the uninformative style of a randomly selected natural image. The outputs are medically irrelevant and resemble the outputs using the Artistic Paintings style source.

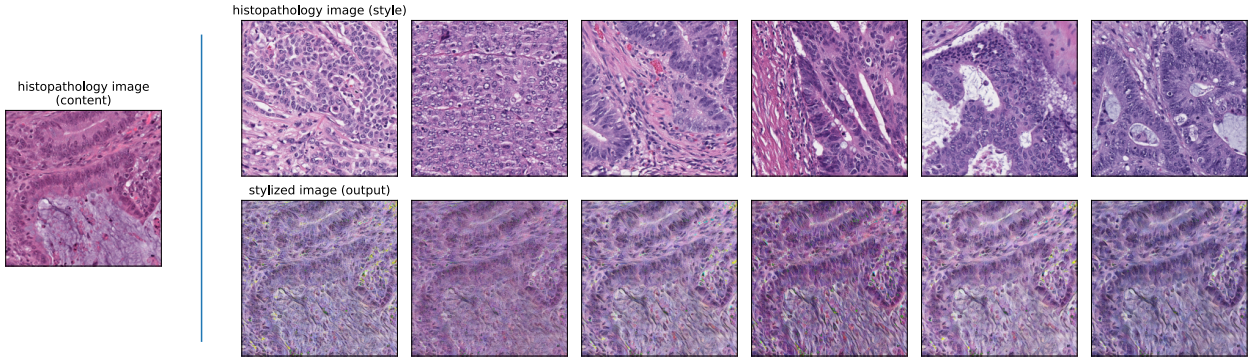

Figure 4: Style transfer with randomly selected histopathologic images from the non-stain normalized version of the Stanford-CRC dataset, applied to a histopathology image (content on the left). The outputs are medically relevant and resemble the outputs obtained with stain augmentation (Figure 5).

We constructed stylized version of the datasets by applying AdaIn style transfer [6] following the method proposed in [4]. AdaIn style transfer takes a content image and an arbitrary style image as inputs, and synthesizes an output image that recombines the content of the former and the style of the latter. After encoding the content and style images in feature space via an encoder, both feature maps are fed to an AdaIN layer that aligns the mean and variance of the content feature maps to those of the style feature maps, producing the target feature maps. Then the stylized output image is generated by a decoder from the target feature maps. We chose AdaIN style transfer because it enables to transfer arbitrary styles in real-time. Each histopathology image was stylized with the style of a randomly selected image from the style source through AdaIN with a stylization coefficient of 1.0. We studied three distinct style sources: 1) artistic paintings from the Kaggle's `Painter by Numbers` dataset ($79,433$ paintings), downloaded via `https://www.kaggle.com/c/painter-by-numbers`, hereafter referred to as the Artistic Paintings style source; 2) natural images from the `miniImageNet` dataset proposed by Vinyals *et al*. [13], consisting of $60,000$ color images from ImageNet with 100 classes, each having 600 examples, hereafter referred to as the Natural Imaging style source; and 3) the original Stanford-CRC dataset, containing $66,578$ histopathological images without stain normalization (to preserve the original variability in staining) as described in section 2.2.1, hereafter referred to as Histopathology Imaging style source. The former two apply medically-irrelevant transformation (Figs. 2 and 3), whereas the latter applies medically-relevant transformation (Fig. 4). Of note, when applying STRAP, we resized the content histopathology images to 1024×1024 pixels and the style source images to 256×256 pixels to maintain geometric features of the content images during the stylization. We prepared stylized version of the datasets in advance, because random style transfer via AdaIN as an on-the-fly data augmentation is still computationally expensive.

We compared STRAP against two standard baseline approaches; stain normalization (SN) and stain augmentation (SA). The STRAP model was trained on stylized datasets alone, whereas the SN model was trained on non-stylized original datasets that were stain-normalized by the Macenko's method [8] and the SA model was trained on non-stylized original datasets with on-the-fly stain augmentation by following the method described by Tellez *et al*. [9] (Fig. 5).

Stain normalization is a widely used method in computational pathology to account for variations in H&E staining [14, 15, 16, 17]. On the other hand, Tellez *et al*. [9] demonstrated that stain augmentation improved classification performance when compared to stain normalization, by increasing the CNN's ability to generalize to unseen stain variations.

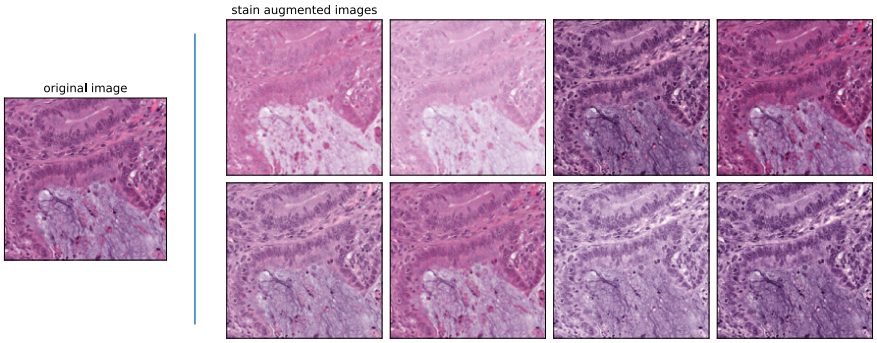

Figure 5: Stain augmentation applied to a histopathology image (original on the left).

2.2 Experiments

We evaluated our proposed approach on two classification tasks, genetic subtype classification and tumor identification, in different domain generalization scenarios, single-domain generalization and multi-source domain generalization, respectively.

2.2.1 Genetic subtype classification in single-domain generalization setting

The genetic subtype classification task was to classify colorectal cancer into two distinct genetic subtypes based on microsatellite status (either microsatellite stable (MSS) or microsatellite unstable (MSI)) using hematoxylin and eosin (H&E)-stained, formalin-fixed, paraffin embedded (FFPE) whole-slide images (WSIs) of surgically resected colorectal cancers. We evaluated our proposed approach in single-domain generalization setting (models are trained on a single-domain dataset and tested on a mixed-domain dataset).

We reused three datasets that were created and used in previous publications; Stanford-CRC from Yamashita *et al*. [16] and CRC-DX-TRAIN as well as CRC-DX-TEST from Kather *et al*. [18] (See the original publications for details, such as inclusion/exclusion criteria and clinico-pathological parameters). These datasets consists of image patches called tiles, which were generated from the WSIs with a size of 512×512 pixels at a resolution of 0.5 µm/pixel and subsequently stain normalized with the Macenko's method [8].

The Stanford-CRC is an in-house dataset that originates from a single institution and contains $66,578$ image tiles ($31,789$ tiles from 50 MSS and $34,789$ tiles from 50 MSI H&E-stained FFPE WSI) from 100 unique patients. The WSIs were originally scanned at $40\times$ base magnification level (0.25 µm/pixel). This single-institutional dataset has equal class distribution, with 50 MSS and 50 MSI patients.

The CRC-DX-TRAIN dataset stems from the TCGA-COAD and TCGA-READ diagnostic slide collections of the Cancer Genome Atlas (TCGA) [19], consisting of data from 18 institutions with various scanners and protocols, *i.e*., a multi-domain dataset, and contains $93,408$ image tiles ($46,704$ tiles from 223 MSS and $46,704$ tiles from 40 MSI H&E-stained FFPE WSI) from 263 unique patients. The WSI were scanned at either $20\times$ or $40\times$ base magnification (0.5 or 0.25 µm/pixel). This multi-institutional dataset was balanced in class distribution.

The CRC-DX-TEST dataset stems from the same diagnostic slide collections of TCGA as the CRC-DX-TRAIN dataset, *i.e*., consisting of data from 18 institutions with various scanners and protocols, and contains $99,904$ image tiles ($70,569$ tiles from 74 MSS and $29,335$ tiles from 26 MSI H&E-stained FFPE WSI) from 100 unique patients. This multi-institutional dataset maintains the original class imbalance, which reflects real-world prevalence of MSI in colorectal cancer.

We performed both in-distribution and out-of-distribution experiments using the above three datasets. For in-distribution analysis, models were trained on CRC-DX-TRAIN and evaluated on CRC-DX-TEST. Our out-of-distribution experiment follows the single-domain generalization setting, where models were trained on single-domain Stanford-CRC dataset and evaluated on multi-source domain CRC-DX-TEST dataset. We applied 4-fold cross-validation to account for the

selection bias introduced by randomness in splitting Stanford-CRC, given its relatively limited sample size; therefore, average and standard deviation of the evaluation metric across the folds were reported. Of note, all the STRAP models were trained on the stylized version of the training datasets by applying the style transfer method described in section 2.1.

We employed the MobileNetv2 [20] model pretrained on ImageNet [21] via transfer learning with stochastic gradient descent with momentum [22], using a fixed learning rate of $4e{-}3$ and epoch of 40, along with early stopping with a patience of five. We used a binary cross entropy loss. All input images were resized to 224×224 pixels before being fed into the network. Random horizontal and vertical flipping (with a probability of 0.5 for each) and random resized cropping were applied as a common data augmentation method. Tile-wise model outputs were aggregated into a patient-wise score by taking their average. The particular metric of interest was the area under the receiver-operating-characteristic curve (AUROC).

We further compared the STRAP model against two state-of-the-arts, Kather *et al*. [15] and Yamashita *et al*. [16] in the same single-domain generalization scenario for genetic subtype classification. Both approaches are similar to our SN baseline, though there are some differences in model architecture, training protocols, and configuration of data augmentation. For example, Kather *et al*. used a ResNet-18 architecture [23] and applied horizontal and vertical flips and random translation along the x and y axes for data augmentation. Similarly, Yamashita *et al*. used a MobileNetV2 architecture and applied data augmentation with random horizontal flips, random rotations, and random color jitter. Model performance for Kather *et al*. [15] and Yamashita *et al*. [16] was either computed using the code available at `https://github.com/jnkather/MSIfromHE` and `https://github.com/rikiyay/MSINet`, respectively, or obtained from the literature.

Impact of differences in style source and stylization coefficient As sensitivity analyses, we performed two additional experiments. First, we studied the effect of difference between medically-irrelevant and medically-relevant STRAP approaches. More specifically, we compared the performance of the STRAP models using Artistic Paintings and Natural Imaging style sources (medically-irrelevant approach) against the STRAP model with Histopathologic Imaging style source (medically-relevant approach) on the genetic subtype classification task. We also studied the effect of difference in stylization coefficient, where the STRAP models using stylization coefficient of 1.0, 0.8, and 0.6 were compared on the genetic subtype classification task.

Assessment on stylized images with random test-time styles To understand how content and style are being utilized, we compared the model performance for STRAP, SA, and SN on stylized version of the CRC-DX-TEST with random test-time styles of the Natural Imaging style source. For STRAP, we tested both STRAP with Artistic Paintings and STRAP with Histopathologic Imaging to assess the difference in sensitivity between medically-irrelevant and medically-relevant approaches.

Assessment on low-frequency components To gain insights into what frequency components the three models (STRAP, SA, and SN) exploit for learning representations, we tested model performance on the low-frequency components of the CRC-DX-TEST dataset, hereafter referred to as LF-CRC-DX-TEST. We constructed the LF-CRC-DX-TEST dataset by following the method described in [11], where all image tiles in the CRC-DX-TEST dataset were decomposed into low- and high-frequency components by applying the fast Fourier transform (FFT) algorithm. Low-frequency components were obtained from the centralized frequency spectrum by applying circular low-pass filters with various radii. All frequencies outside the circular filter were set to zero and the inverse FFT was applied subsequently to get the low-frequency images (Fig. 6). To identify the low-pass filter size that corresponds to the highest model performance, the AUROC for each of the STRAP, SA, and SN models was assessed using varying low-pass filter sizes (the radii ranged from 14 to 154).

We also visualized saliency maps on the LF-CRC-DX-TEST (with a low-pass filter size of 70) using integrated gradients attributions [10] to highlight which pixels of an input image contribute more to model inference.

2.2.2 Tumor identification in multi-domain generalization setting

The tumor identification task was to classify presence or absence of breast cancer metastases in image patches extracted from histopathlogic scans of lymph node sections in multi-source domain generalization setting (models are trained on a multi-source domain dataset and tested on a single-domain dataset).

We used the CAMELYON17-WILDS dataset [24], a patch-based variant of the original Camelyon17 dataset [25] created as a benchmark dataset for domain generalization, where the domains are hospitals and the goal is to learn models that generalize to data from a hospital that is not in the training subset. The specific task is to predict if a given region of tissue contains any tumor tissue, which was modeled as binary classification, where the input is a

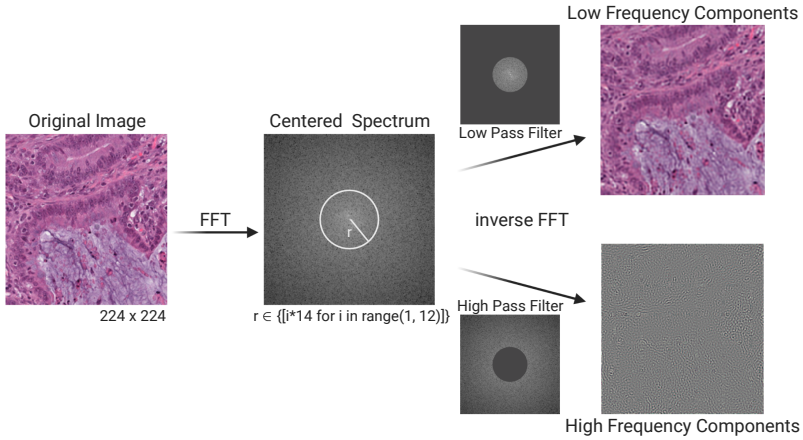

Figure 6: A schema for generating low-frequency components of an image. Image tiles are decomposed into low- and high-frequency components by applying the fast Fourier transform (FFT) algorithm. Low-frequency components are extracted from the centralized frequency spectrum by applying circular low-pass filters with various radii. All frequencies outside the circle were set to zero and the inverse FFT was applied subsequently. Of note, the high frequency components were not used in this study.

96×96-pexel histopathological image, the label is a binary indicator of whether the central 32×32 region contains any tumor tissue.

The CAMELYON17-WILDS dataset was adapted from WSIs of breast cancer metastases in lymph nodes sections, obtained from the CAMELYON17 challenge [25], where the WSIs were scanned at a resolution of 0.23–$0.25\mu\text{m}$, and each WSI contains multiple resolution levels, with approximately $10,000 \times 20,000$ pixels at the highest resolution level. Image patches were generated using the third-highest resolution level, corresponding to reducing the size of each dimension by a factor of 4. The CAMELYON17-WILDS dataset comprises $455,954$ patches extracted from 50 WSIs of breast cancer metastases in lymph node sections, with 10 WSIs from each of 5 hospitals. The label for each patch was determined by the segmentation masks manually annotated with tumor regions by pathologists, which were provided along with the original Camelyon17 dataset. We split the CAMELYON17-WILDS dataset by domain (*i.e.*, which hospital the patches were taken from) using the metadata. We used the `Test(OOD)` subset of the CAMELYON17-WILDS dataset as our out-of-distribution test subset, which contains $85,054$ patches taken from 10 WSIs from the 5th hospital (`center` 2 in the provided metadata), which was chosen by the original WILDS project because its patches were the most visually distinctive. We split the rest patches into training and validation based on the `split` column provided in the metadata (`split` 0 for training and `split` 1 for validation), where $333,866$ and $37,034$ patches taken from 40 WSIs, with 10 WSIs from each of the 4 hospitals, were assigned to the training and validation sets, respectively. Of note, the training/validation and test sets comprise class-balanced patches from separate hospitals (See the original publication [24] for more details).

We employed the ResNet-50 [23] model pretrained on ImageNet [21] via transfer learning with stochastic gradient descent with momentum [22], using a fixed learning rate of $4\text{e}{-3}$ and epoch of 40, along with early stopping with a patience of five. We used a binary cross entropy loss. Random horizontal and vertical flipping (with a probability of 0.5 for each) and random resized cropping were applied as a common data augmentation method. Model performance was evaluated by average accuracy and AUROC across patches. Of note, unlike the genetic subtype classification task where the ground truth labels are patient-level, the ground truth labels for the tumor identification task are patch-level, meaning no output aggregation procedure is required.

2.3 Statistical analysis

We assessed model performance using the AUROC for genetic subtype classification, and accuracy as well as AUROC for tumor identification. We calculated 95% confidence intervals (CI) using bootstrapping with the percentile method with $2,000$ resamples. Statistical comparisons were performed using the DeLong's test [26] for individual AUROC, a paired t-test for average AUROC, and a permutation test with $2,000$ resamples for accuracy. For the main analyses of both genetic subtype classification and tumor identification (results are shown in Tables 1 and 5, respectively), p-values were adjusted using the Benjamini-Hochberg method [27] to account for multiple comparisons by controlling the false positive rate to less than 0.10. Otherwise, a two-tailed alpha criterion of 0.05 was used for statistical significance.

3 Experimental Results

3.1 Genetic subtype classification in single-domain generalization setting

The STRAP model with Artistic Paintings style source achieved an average AUROC of 0.876 on the out-of-distribution multi-domain CRC-DX-TEST dataset, and outperformed the SA, SN, and the two state-of-the-art models (Table 1). STRAP also demonstrated a minimal, even negative, performance drop from in-distribution to out-of-distribution testing (see column `Delta` in Table 1), whereas SA presented near-zero performance drop and the others showed positive performance drops. These results suggest that the STRAP model has the ability to learn more discriminative and generalizable (*i.e.*, class-specific and domain-irrelevant) visual representations, compared to the other approaches that may exploit some extent of the domain-specific features.

Table 1: Comparison of style transfer augmentation (STRAP), stain augmentation (SA), stain normalization (SN), and two state-of-the-arts on in-distribution and out-of-distribution (single-domain generalization) scenarios on the genetic subtype classification task.

	CRC-DX-TRAIN → CRC-DX-TEST (ID)		Stanford-CRC → CRC-DX-TEST (OOD)		Delta§
	AUROC†	p-value (vs STRAP)	AUROC‡	p-value (vs STRAP)	(ID−OOD)
STRAP (AP)	**0.847 [0.741, 0.932]**	REF	**0.876 (0.015)**	REF	**−0.029**
SA	0.816 [0.709, 0.917]	0.471	0.814 (0.020)	0.002∗	0.002
SN	0.794 [0.684, 0.892]	0.456	0.765 (0.031)	0.003∗	0.029
Kather *et al.*	0.759 [0.632, 0.873]	0.219	0.742 (0.013)	0.001∗	0.018
Yamashita *et al.*	0.816 [0.712, 0.914]	0.456	0.786 (0.020)	0.010∗	0.030

Arrows indicate: train data → test data, *e.g.*, CRC-DX-TRAIN → CRC-DX-TEST means training on CRC-DX-TRAIN and testing on CRC-DX-TEST.
∗ indicates a significant difference.
† represents AUROC with 95% CI in square brackets.
‡ represents average AUROC of models obtained via cross-validation, with standard deviation in parentheses.
§ indicates average performance drop from in-distribution (CRC-DX-TRAIN → CRC-DX-TEST) to out-of-distribution (Stanford-CRC → CRC-DX-TEST) scenarios.
Stylization coefficient (alpha) of 1.0 was used for the STRAP model.
P-values were adjusted using the Benjamini-Hochberg method [27].
Abbreviations: AP, Artistic Paintings; AUROC, areas under the receiver-operating-characteristic curve; CV, cross-validation; ID, in-distribution; OOD, out-of-distribution; SA, stain augmentation; SN, style normalization; STRAP, style transfer augmentation.

3.1.1 Impact of differences in style source

For genetic subtype classification, medically-irrelevant STRAP using Artistic Paintings and Natural Imaging as style sources achieved superior performance compared to the medically-relevant STRAP using Histopathologic Imaging as style source. In comparison to Histopathologic Imaging, the Artistic Paintingsyielded a significantly higher performance, whereas there was no statistically significant difference between the Natural Imaging and Histopathologic Imaging style sources (Table 2).

3.1.2 Impact of stylization coefficient

We also tested the effect of the stylization coefficient on STRAP model performance. We found that, among stylization coefficients of 1.0, 0.8, and 0.6, the larger the stylization coefficient (*i.e.*, with a stylization coefficient of 1.0), the higher the model performance (Table 3), which suggests that the STRAP model can learn more discriminative and generalizable representations when more low-level content within an image was removed and replaced by the style transfer operation.

3.1.3 Assessment on stylized images with random test-time styles

We assessed the model performance on stylized version of CRC-DX-TEST created using Natural Imaging as style source. As shown in Table 4, STRAP with Artistic Paintings style source, a medically-irrelevant style transfer, achieved significantly higher performance compared to SA and SN and tended to have higher performance compared to medically-relevant STRAP with Histopathologic Imaging style source. STRAP with Artistic Paintings also demonstrated the smallest performance difference between original and stylized CRC-DX-TEST. This result suggests that the medically-irrelevant STRAP successfully biased the networks to content/shape, which may explain its superior performance

Table 2: Effect of different style sources on STRAP model performance.

Style Source	Stanford-CRC → CRC-DX-TEST AUROC†	Stanford-CRC → CRC-DX-TEST p-value (vs HI)
Artistic Paintings (AP)	**0.876 (0.015)**	0.037∗
Natural Imaging (NI)	0.867 (0.016)	0.088
Histopathologic Imaging (HI)	0.822 (0.042)	REF

Arrow indicates: train data → test data, *i.e.*, Stanford-CRC → CRC-DX-TEST means training on Stanford-CRC and testing on CRC-DX-TEST.
∗ indicates a significant difference.
† represents average AUROC of models obtained via cross-validation, with standard deviation in parentheses.
Stylization coefficient (alpha) of 1.0 was used for the STRAP model.
Abbreviations: AUROC, areas under the receiver-operating-characteristic curve; CV, cross-validation.

Table 3: Effect of stylization coefficient on STRAP model performance.

Stylization Coefficient	Stanford-CRC → CRC-DX-TEST AUROC†	Stanford-CRC → CRC-DX-TEST p-value (vs SC 1.0)
SC 1.0	**0.876 (0.015)**	REF
SC 0.8	0.856 (0.036)	0.189
SC 0.6	0.846 (0.024)	0.024∗

Arrow indicates: train data → test data, *i.e.*, Stanford-CRC → CRC-DX-TEST means training on Stanford-CRC and testing on CRC-DX-TEST.
∗ indicates a significant difference.
† represents average AUROC of models obtained via cross-validation, with standard deviation in parentheses.
Abbreviations: AUROC, areas under the receiver-operating-characteristic curve; CV, cross-validation; SC, stylization coefficient.

and out-of-distribution generalizability compared to the other three (*i.e.*, medically-relevant STRAP, SA, and SN) approaches.

3.1.4 Assessment on low-frequency components

We evaluated the STRAP, SA, and SN models on the LF-CRC-DX-TEST dataset with a wide range of low-pass filter sizes. As shown in Fig. 7, the STRAP model reached its peak performance at a radius of 84, whereas the other two reached their peaks at a radius of 112. These results suggest that the STRAP model can exploit lower-frequency components for learning representations, whereas the other two baselines rely more on higher-frequency components. We speculate that, because the STRAP model is biased toward shape [4], it performs well on lower-frequency components, which preserve most of the geometry and thus, almost look identical to the original image to human. On the contrary, the baseline SA and SN approaches do not address style and content explicitly and thus, require texture and/or higher-frequency components to reach their peak performance.

Saliency maps with integrated gradients show that the STRAP model presented high attributions at specific areas and less diffusely distributed attributions, whereas the SA and SN models showed more broadly distributed attributions that might correspond to the low-level texture content of the images (Fig. 8). A board-certified, subspecialty gastrointestinal pathologist interpreted these saliency maps and concluded that STRAP picks up tumor-infiltrating lymphocytes as well as mitotic figures, which are well-known human-recognizable histomorphologic features that are associated with the genetic subtype of interest.

Table 4: Model performance on CRC-DX-TEST dataset with and without random test-time styles.

	AUROC on CRC-DX-TEST†			
	Original	Stylized	Delta§	p-value‡
STRAP (AP)	**0.876 (0.015)**	**0.830 (0.020)**	**0.046**	REF
STRAP (HI)	0.822 (0.042)	0.711 (0.077)	0.111	0.085
SA	0.814 (0.020)	0.726 (0.055)	0.084	0.047∗
SN	0.765 (0.031)	0.633 (0.577)	0.132	0.015∗

∗ indicates a significant difference.
† represents average AUROC with standard deviation in parentheses.
§ represents average performance difference between original and stylized CRC-DX-TEST.
‡ represents p-value for comparing model perfromance on stylized CRC-DX-TEST.
Stylized CRC-DX-TEST was created using Natural Imaging style source.
Stylization coefficient (alpha) of 1.0 was used for the STRAP models.
Abbreviations: AP, Artistic Paintings; AUROC, areas under the receiver-operating-characteristic curve; HI, Histopathologic Imaging; SA, stain augmentation; SN, style normalization; STRAP, style transfer augmentation.

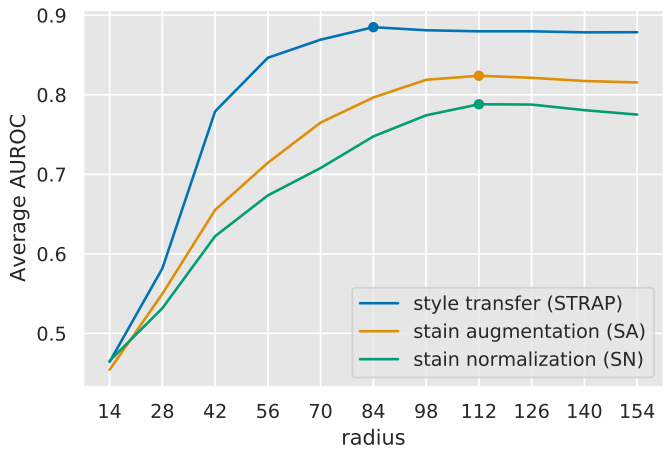

Figure 7: Results of the experiments using the low-frequency components of the CRC-DX-TEST dataset (LF-CRC-DX-TEST). The x-axis represents the radii of low-pass filters used to generate the LF-CRC-DX-TEST dataset, and the y-axis shows the average area under the receiver-operating-characteristic curves (AUROC) across cross-validation folds. Each dot marker represents the corresponding peak performance.

3.2 Tumor identification in multi-domain generalization setting

On the tumor identification task using CAMELYON17-WILDS dataset, we developed models on the tiles from four out of five hospitals, and assessed the performance on the tiles from the 5th hospital, *i.e.*, the `Test(OOD)` subset of the CAMELYON17-WILDS (multi-domain generalization setting). As shown in Table 5, the medically-irrelevant STRAP model using Artistic Paintings style source achieved the highest accuracy and AUROC with significant differences compared to the other approaches. The results have a similar trend as those for genetic subtype classification, where another medically-irrelevant STRAP with Natural Imaging style source demonstrated the second highest performance, medically-relevant STRAP using Histopathologic Imaging style source and SA presented similar performance that was the next highest, and SN showed the lowest performance.

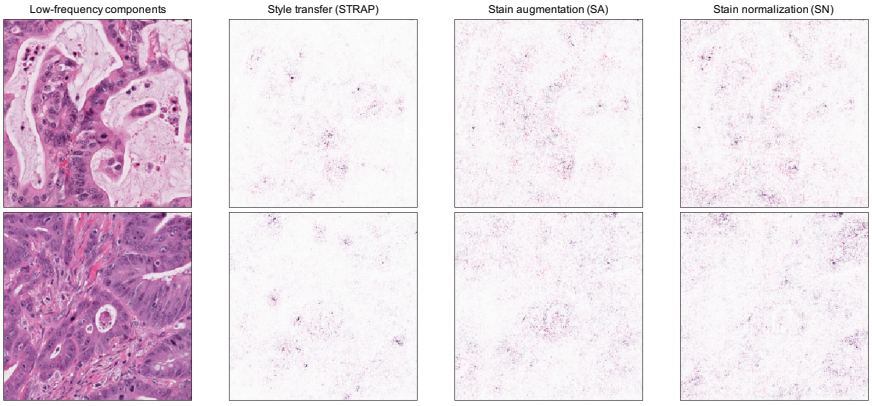

Figure 8: Pixel-wise integrated gradient attributions of the low-frequency components (generated with a radius of 70) of the CRC-DX-TEST dataset (LF-CRC-DX-TEST), visualized as saliency maps for the STRAP, SA, and SN models.

Table 5: Comparison of style transfer augmentation (STRAP), stain augmentation (SA), stain normalization (SN) on out-of-distribution (multi-source domain generalization) scenarios on the tumor identification task.

	CAMELYON17-WILDS			
	Accuracy	p-value	AUROC	p-value
STRAP (AP)	**0.937 [0.935, 0.938]**	REF	**0.981 [0.980, 0.982]**	REF
STRAP (NI)	0.923 [0.921, 0.925]	<0.0001∗	0.977 [0.976, 0.978]	<0.0001∗
STRAP (HI)	0.831 [0.829, 0.834]	<0.0001∗	0.888 [0.885, 0.890]	<0.0001∗
SA	0.833 [0.830, 0.835]	<0.0001∗	0.916 [0.914, 0.918]	<0.0001∗
SN	0.631 [0.628, 0.634]	<0.0001∗	0.859 [0.856, 0.861]	<0.0001∗

∗ indicates a significant difference.
Stylization coefficient (alpha) of 1.0 was used for the STRAP models.
P-values were adjusted using the Benjamini-Hochberg method [27].
Abbreviations: AP, Artistic Paintings; AUROC, areas under the receiver-operating-characteristic curve; HI, Histopathologic Imaging; NI, Natural Imaging; SA, stain augmentation; SN, style normalization; STRAP, style transfer augmentation.

4 Discussion

We present **STRAP** (**S**tyle **TR**ansfer **A**ugmentation for histo**P**athology), which achieved improved performance and generalizability when compared with two standard baselines (stain augmentation (SA) and stain normalization (SN)) on two classification tasks (*i.e.*, genetic subtype classification in single-domain generalization setting, and tumor identification in multi-domain generalization setting) using digitized histopathology images in computational pathology.

We speculate that STRAP helps models learn domain-agnostic and class-specific visual representations by removing the original texture and/or high-frequency components from the histopathology images, which are domain-specific and class-irrelevant, and predominantly leaving shape-biased and/or low-frequency content, which are domain-irrelevant and class-specific. In fact, more intensive style transfer with a higher stylization coefficient resulted in superior performance. Furthermore, when tested on stylized version of the out-of-distribution test dataset with random test-time styles, STRAP with Artistic Paintings showed significantly higher performance compared to the baseline SA and SN approaches. Also, our experiments on the low-frequency components demonstrated that the STRAP approach helps models exploit lower frequency components, in contrast to the standard SA and SN approaches that rely more on higher frequency components. This speculation is also consistent with the hypotheses proposed by Geirhos *et al*. [4] and Wang *et al*. [11]—that shape-biased and/or low-frequency features are essential for deep learning models to learn robust and generalizable visual representations.

To the best of our knowledge, no previous study has applied medically-irrelevant image manipulation for the development of deep learning models for medical imaging. Four previous studies have applied the style transfer technique to medical imaging tasks in computational pathology [28, 29] and skin lesion classification [30, 31]. However, these studies employed medically-relevant transformation with the aim of combating data scarcity, class imbalance, and stain

variation. Our study demonstrates that medically-irrelevant transformation, *i.e*., STRAP with Artistic Paintings or Natural Imaging style sources, can result in improved performance and generalizability, when compared with medically-relevant transformation, *i.e*., style transfer with Histopathologic Imaging style source and stain augmentation. A possible explanation for this phenomenon is that medically-irrelevant style transfer can result in a wider variety of transformation using a more diverse set of styles compared to the medically-relevant approaches for which the variations in color and texture are more uniform and thus, limited. Tobin *et al*. [32] showed that an object detection model that generalizes to real-world images can be trained by using unrealistic simulated images with a diverse set of random textures, rather than by making the simulated images as realistic as possible. As in the human learning process, learning class-specific and domain-irrelevant patterns from data is essential for deep learning models, and the style transfer technique with a diverse set of random styles can be a powerful tool to control the representations models learn.

Although data augmentation is widely used when training deep learning models for medical imaging tasks, its potential has not yet been fully studied and still remains an active area of research. Moreover, an optimal configuration of data augmentation methods may vary among applications. As our study suggests, data augmentation can be a simple yet powerful tool for learning domain-agnostic representation. Further research is warranted to identify optimal data augmentation techniques for a variety of medical imaging tasks, and medically-irrelevant transformations such as the proposed STRAP approach should be considered, along with established methods.

As shown in Table 1, STRAP with Artistic Paintings achieved higher performance in the out-of-distribution setting, compared to the in-distribution setting, whereas opposite results were observed for the other baseline approaches and state-of-the-arts. As described in Section 2.2.1, the training data in the in-distribution setting was a multi-source domain dataset, whereas the training data used for the out-of-distribution setting was a single-source dataset. Although it is often said that diverse multi-institutional datasets are needed for training models that generalize on unseen data [33], our study may suggest that a well-curated homogeneous dataset could provide value in training domain-agnostic models, if a model has sufficient capability to learn domain-invariant and class-specific representations, similar to the way in which humans learn from a set of representative examples (*e.g*., content presented in textbooks).

Besides supervised learning, our approach may be applicable to self-supervised learning. A contrastive learning framework, such as SimCLR [34] and MoCo [35], learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss (thus, relying heavily on a stochastic data augmentation module). Chen *et al*. [34] showed that the composition of data augmentation operations is crucial in yielding effective representations, and that unsupervised contrastive learning benefits from strong data augmentation. In medical imaging, contrastive learning may require a tailored composition of data augmentation operations, and our medically-irrelevant STRAP has the potential to serve as one of the core transformation operations.

One limitation of this study is that we only tested our approach with classification tasks in the field of computational pathology. Further studies are warranted to investigate whether our approach could prove its efficacy and robustness 1) for non-classification tasks such as detection, segmentation, and survival prediction, and 2) in other medical imaging domains, such as radiology, ophthalmology, and dermatology. Another limitation is STRAP's relatively longer runtime compared to the other two baseline approaches, where the average runtime was 1.08 s for STRAP, 8.13 ms for SA, and 6.42 ms for SN on a workstation with a GeForce RTX 2080 Ti (NVIDIA, Santa Clara, CA) graphics processing unit, a Core i9-9820X (10 cores, 3·3 GHz) central processing unit (Intel, Santa Clara, CA, and 128 GB of random-access memory. Improvement in computational efficiency is required to apply STRAP as one of on-the-fly data augmentations.

In conclusion, we have introduced STRAP, a form of data augmentation based on random style transfer with medically-irrelevant style source, for learning domain-agnostic visual representations in computational pathology. Our experiments demonstrated that our approach yields significant improvements in test performance on classification tasks in computational pathology, particularly in the presence of domain shift. Our study provides evidence that 1) CNNs are reliant on low-level texture content and are therefore vulnerable to domain shifts in computational pathology, and that 2) medically-irrelevant STRAP can be a practical tool for mitigating that reliance and, therefore, a possible solution for learning domain-agnostic representations.

Acknowledgements

This work was funded by the Stanford Departments of Biomedical Data Science and Pathology, through a Stanford Clinical Data Science Fellowship to RY. We would like to thank Blaine Burton Rister for detailed and valuable feedback on the manuscript. We would also like to thank Nandita Bhaskhar, Khaled Kamal Saab, and Jared Dunnmon for their helpful discussions.

References

[1] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset Shift in Machine Learning*. MIT Press, 2009.

[2] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *ICML*, 2019.

[3] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, June 2016.

[6] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[7] Philip T G Jackson, Amir Atapour Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: data augmentation via style randomization. In *CVPR Workshops*, pages 83–92, 2019.

[8] Marc Macenko, Marc Niethammer, J S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. ieeexplore.ieee.org, June 2009.

[9] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.*, 58:101544, December 2019.

[10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[11] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020.

[12] Bhuvanesh Awasthi, Jason Friedman, and Mark A Williams. Faster, stronger, lateralized: low spatial frequency information supports face processing. *Neuropsychologia*, 49(13):3583–3590, November 2011.

[13] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D Lee, M Sugiyama, U Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3630–3638. Curran Associates, Inc., 2016.

[14] Amelie Echle, Heike Irmgard Grabsch, Philip Quirke, Piet A van den Brandt, Nicholas P West, Gordon G A Hutchins, Lara R Heij, Xiuxiang Tan, Susan D Richman, Jeremias Krause, Elizabeth Alwers, Josien Jenniskens, Kelly Offermans, Richard Gray, Hermann Brenner, Jenny Chang-Claude, Christian Trautwein, Alexander T Pearson, Peter Boor, Tom Luedde, Nadine Therese Gaisa, Michael Hoffmeister, and Jakob Nikolas Kather. Clinical-Grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*, 159(4):1406–1416.e11, October 2020.

[15] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, Heike I Grabsch, Takaki Yoshikawa, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, Christian Trautwein, and Tom Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.*, 25(7):1054–1056, June 2019.

[16] Rikiya Yamashita, Jin Long, Teri Longacre, Lan Peng, Gerald Berry, Brock Martin, John Higgins, Daniel L Rubin, and Jeanne Shen. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol.*, 22(1):132–141, January 2021.

[17] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Comput. Struct. Biotechnol. J.*, 16:34–42, February 2018.

[18] Jakob Nikolas Kather. Histological images for MSI vs. MSS classification in gastrointestinal cancer, FFPE samples [data set] Zenodo. http://doi.org/10.5281/zenodo.2530835, 2019.

[19] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, July 2012.

[20] M Sandler, A Howard, M Zhu, A Zhmoginov, and L Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, June 2018.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, December 2015.

[22] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Netw.*, 12(1):145–151, January 1999.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[24] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

[25] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

[26] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[27] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, 1995.

[28] Pietro Antonio Cicalese, Aryan Mobiny, Pengyu Yuan, Jan Becker, Chandra Mohan, and Hien Van Nguyen. StyPath: Style-Transfer data augmentation for robust histology image classification. *arXiv preprint arXiv:2007.05008*, 2020.

[29] Seo Jeong Shin, Seng Chan You, Hokyun Jeon, Ji Won Jung, Min Ho An, Rae Woong Park, and Jin Roh. Style transfer strategy for developing a generalizable deep learning application in digital pathology. *Comput. Methods Programs Biomed.*, 198:105815, January 2021.

[30] Agnieszka Mikołajczyk and Michał Grochowski. Style transfer-based image synthesis as an efficient regularization technique in deep learning. In *2019 24th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pages 42–47, 2019.

[31] Tamás Nyíri and Attila Kiss. Style transfer for dermatological data augmentation. In *Intelligent Systems and Applications*, pages 915–923. Springer International Publishing, 2020.

[32] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *arXiv preprint arXiv:1703.06907*, 2017.

[33] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.*, 17(1):195, October 2019.

[34] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[35] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.