Self-supervised Vision Transformer are Scalable Generative Models for Domain Generalization

Sebastian Doerrich, Francesco Di Salvo, and Christian Ledig

xAILab Bamberg, University of Bamberg, Germany
sebastian.doerrich@uni-bamberg.de

Abstract. Despite notable advancements, the integration of deep learning (DL) techniques into impactful clinical applications, particularly in the realm of digital histopathology, has been hindered by challenges associated with achieving robust generalization across diverse imaging domains and characteristics. Traditional mitigation strategies in this field such as data augmentation and stain color normalization have proven insufficient in addressing this limitation, necessitating the exploration of alternative methodologies. To this end, we propose a novel generative method for domain generalization in histopathology images. Our method employs a generative, self-supervised Vision Transformer to dynamically extract characteristics of image patches and seamlessly infuse them into the original images, thereby creating novel, synthetic images with diverse attributes. By enriching the dataset with such synthesized images, we aim to enhance its holistic nature, facilitating improved generalization of DL models to unseen domains. Extensive experiments conducted on two distinct histopathology datasets demonstrate the effectiveness of our proposed approach, outperforming the state of the art substantially, on the CAMELYON17-WILDS challenge dataset ($+2\%$) and on a second epithelium-stroma dataset ($+26\%$). Furthermore, we emphasize our method's ability to readily scale with increasingly available unlabeled data samples and more complex, higher parametric architectures. Source code is available at github.com/sdoerrich97/vits-are-generative-models .

Keywords: domain generalization · self-supervised learning · feature orthogonalization · generative image synthesis.

1 Introduction

Deep learning (DL) has had a significant impact on a broad range of domains ranging from image classification to natural language processing [24]. Nevertheless, its incorporation into routinely used medical image analysis has progressed comparatively slow [21], mainly due to difficulties in achieving robust generalization across diverse imaging domains. This challenge is particularly pronounced in digital histopathology, where variations in coloring agents and staining protocols for histological specimens exacerbate domain disparity [16]. Traditional approaches to address these generalizability challenges in digital histopathology typically involve data augmentation or stain color normalization [2]. Data
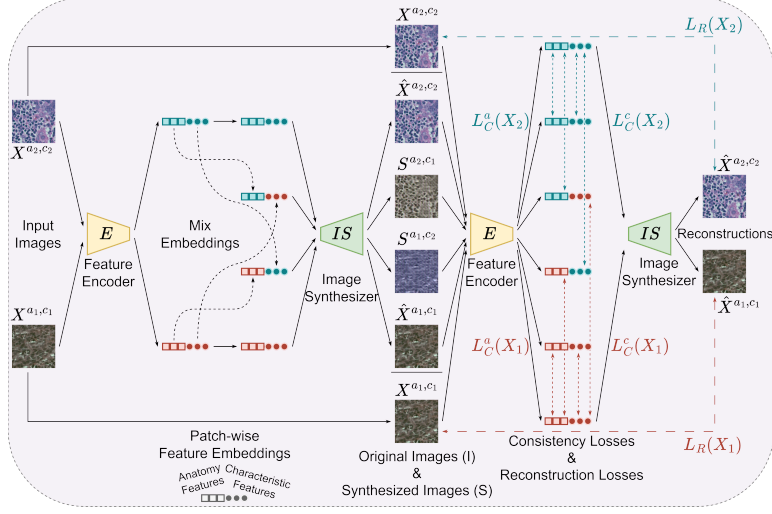
Fig. 1: Schematic Visualization of our self-supervised generative approach. A single ViT encoder (E) is used to separate anatomy from image-characteristic features of distinct images which are subsequently intermixed among each other and processed by an image synthesizer (IS) to generate synthetic images.

augmentation techniques manipulate aspects of color [12], apply stain-specific channel-wise augmentation [22], or incorporate stain colors of unseen domains into the training data [2]. Alternatively, stain color normalization aligns images' color patterns using target domain information [18,15,23]. However, these methods often require access to target samples during training or struggle with adapting to new domains and unseen stain colors. To overcome these limitations, Lafarge et al. [10] investigate the use of Domain Adversarial Neural Networks (DANNs) to enhance cross-domain performance. Conversely, Nguyen et al. [17] propose ContriMix, which aims to improve domain generalization by augmenting the diversity of the source domain with synthetic images. This is achieved by initially separating biological content from technical variations and subsequently combining them to form new anatomy-characteristic combinations. However, ContriMix's dependence on convolutional encoders restricts the diversity of its synthetic images, as it allows for the extraction of only a single characteristic tensor per image. In this work, we focus on those limitations and present a novel generative domain generalization (DG) method for histopathology images. Employing a self-supervised Vision Transformer (ViT), we generate synthetic images with diverse combinations of anatomy and image characteristics, enriching the holistic nature of the dataset without requiring any domain information. This allows DL models trained on the extended dataset to adapt to unseen domains more effectively. To prove this, we evaluate our method in extensive experiments against the current state of the art on two distinct benchmark datasets for domain generalization in histopathology.

Our main contributions are:

- We present a novel self-supervised generative domain generalization method for histopathology.
- We generate synthetic images with unseen combinations of anatomy and image characteristics.
- We extensively evaluate our method on two histopathology benchmark datasets and outperform the state of the art by a large margin.
- We assess our method's ability to scale effectively with growing availability of unlabeled data samples and the adoption of deeper architectures.

2   Method

Our method is a self-supervised generative approach that employs feature orthogonalization to generate synthetic images. Using a single ViT encoder (E), we encode an image patch-wise and split the resulting embeddings, with one half preserving anatomy and the other half storing characteristic features for each patch. These feature vectors are then mixed across different input images and fed into an image synthesizer (IS) to create synthetic images representing new anatomy-characteristic pairs. See Fig. 1 for an illustration of this process.

2.1   Feature Orthogonalization and Image Synthesis

Taking inspiration from ViT principles [4], we first partition images x_i with $x_i \in \mathbb{R}^{C \times H \times W}$, where C, H, and W are the number of channels, height, and width of the image, respectively, into non-overlapping patches. This results in $\tilde{x}_i \in \mathbb{R}^{P \times C \times PS \times PS}$, where P denotes the number of patches and PS the patch size. These patches are processed by the encoder E to extract feature embeddings z_i for each image. Let $z_i = E(\tilde{x}_i) \in \mathbb{R}^{P \times L}$, where L denotes the encoder's latent dimension, we extract the anatomical ($z_i^a \in \mathbb{R}^{P \times L/2}$) and characteristic ($z_i^c \in \mathbb{R}^{P \times L/2}$) feature vectors by splitting z_i along L. To reconstruct the original images $\hat{x}_i$, the image synthesizer IS reshapes the feature vectors into matrices $Z_i^a \in \mathbb{R}^{P \times C \times PS \times V}$ and $Z_i^c \in \mathbb{R}^{P \times C \times V \times PS}$, where V is the hidden dimension, before applying the dot-product of both feature matrices along V to restore $\hat{x}_i$.

$$\hat{x}_i = IS(z_i^a, z_i^c) = Z_i^a \cdot Z_i^c, \qquad \text{with } \hat{x}_i \in \mathbb{R}^{P \times C \times PS \times PS} \longleftrightarrow \mathbb{R}^{C \times H \times W} \qquad (1)$$

Conversely, to generate synthetic images s_i with diverse anatomy-characteristics combinations, we combine the anatomical feature embeddings z_i^a of each sample x_i in batch b with M characteristic feature embeddings. These are each extracted from a single patch of another sample x_m within the same batch ($m \in 1, \ldots, M$). This patch, and thereby its corresponding characteristic embedding $z_{m,p}^c$ are chosen uniformly at random from each sample x_m. Note that we do not use the entire z_m^c since using the characteristics of a single patch yields substantially more diverse synthetic images. These combinations (z_i^a, $z_{m,p}^c$) are then passed through IS to create the synthetic images s_i, preserving the original anatomy but with severely altered characteristics. This process enables the extraction of fine-grained characteristics, resulting in a diverse range of synthetic images s_i.

2.2   Feature Consistency and Self-Reconstruction

To guide the feature orthogonalization and synthetic image generation, we employ three distinct mean squared error (MSE) loss terms, namely anatomical consistency L_C^a, characteristic consistency L_C^c and self-reconstruction L_R. The anatomical consistency L_C^a for batch b with N training samples and M number of anatomy-characteristic mixes:

$$L_C^a = \frac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} ||z_i^a - z_s^a||_2^2 \tag{2}$$

$$\text{with} \quad z_i^a = E(x_i)^{P \times [1:L/2]} \quad \text{and} \quad z_s^a = E\left(IS\left(z_i^a, z_{m,p}^c\right)\right)^{P \times [1:L/2]}$$

where $z_{m,p}^c$ being the characteristic embedding of a randomly chosen patch p of sample x_m, promotes consistency between the anatomy extracted from the original images x_i and the corresponding synthetic images s_i. In addition, the characteristic consistency L_C^c for batch b with N training samples and M number of anatomy-characteristic mixes:

$$L_C^c = \frac{1}{NMP} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{q=1}^{P} ||z_{m,p}^c - z_{s,q}^c||_2^2 \tag{3}$$

$$\text{with} \quad z_{s,q}^c = E\left(IS\left(z_i^a, z_{m,p}^c\right)\right) \text{ at patch } q \in P \quad \text{and} \quad z_{s,q}^c \in \mathbb{R}^{1 \times L/2}$$

aligns the characteristics of the synthetic images s_i with the characteristic $z_{m,p}^c$ used to create these synthetic images. Lastly, the self-reconstruction loss L_R:

$$L_R = \frac{1}{N} ||x_i - IS(z_i^a, z_i^c)||_2^2 \tag{4}$$

aims to ensure that the self-reconstructed images closely resemble the original ones. Thereby, the combined loss across a set of mini-batches with $b \in 1, \ldots, B$ can be written as:

$$L = \frac{1}{B} \sum_{b=1}^{B} \lambda_a L_C^a + \lambda_c L_C^c + \lambda_r L_R \tag{5}$$

with $\lambda_a, \lambda_c, \lambda_r$ being weights to adjust the influence of each loss during training.

2.3   Training

The encoder is trained independently for each dataset adhering to the objective described above. This fully self-supervised approach allows us to incorporate labeled or unlabeled samples for the anatomical area of interest and facilitates dynamic transfer to additional tasks without retraining. For the ViT encoder E, we opt for the ViT-B/16 backbone, which operates on 224×224 pixel images, splitting them into 16×16 pixel patches and encoding each patch into a 768-dimensional vector. Following [17], we use 4 mixes (number of combinations M of anatomy and characteristics to get synthetic images) per batch. We set $\lambda_a = \lambda_c = \lambda_r = 1$ and train the encoder for 50 epochs with a batch size of 64, utilizing the AdamW optimizer [14] with a learning rate of 0.001, and a cosine annealing learning rate scheduler [13] with a single cycle.

3   Experiments and Results

We assess the domain generalization ability of our method on two histopathology datasets. The first is the Camelyon17-wilds challenge dataset [9,20], focusing on tumor identification across various hospitals. It comprises 96×96 image patches from lymph node whole-slide images, with labels indicating tumor presence in the central 32×32 region. We use the same training (302,436 samples), validation (34,904), and test (85,054) splits as the original publication [9]. For the second dataset, we aggregate three public histopathology datasets: NKI [1], VGH [1], and IHC [11], focusing on epithelium-stroma classification. The NKI (8,337 samples) and VGH (5,920) datasets comprise H&E stained breast cancer tissue images, while the IHC dataset (1,376) consists of IHC-stained colorectal cancer tissue images. Following [8], we alternate between NKI and VGH as the train/validation set, but maintain IHC as the fixed test set due to its distinct coloration. This allows us to mimic a similar generalization challenge as presented in Camelyon17-wilds, where both the validation and test set comprise out-of-distribution (OOD) samples. In order to fully utilize our ViT encoder's abilities, both benchmark datasets are standardized to 224×224 images using bicubic interpolation. Examples for each dataset are illustrated in Fig. 2.

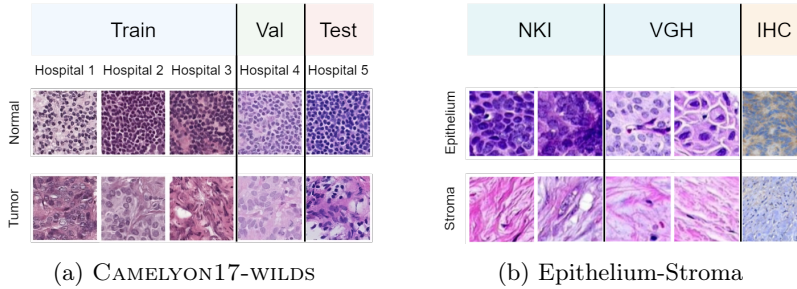(a) Camelyon17-wilds   (b) Epithelium-Stroma

Fig. 2: Examples from the histopathology datasets used for evaluating domain generalization. Left: Camelyon17-wilds for which the domains are hospitals. Right: Combined epithelium-stroma dataset for which the domains are datasets.

3.1   Qualitative Evaluation

We qualitatively evaluate our method by training it on the Camelyon17-wilds dataset and assessing the image quality of the image synthesizer's reconstructions (no mixing). For the training set, we achieve an average Peak Signal-to-Noise Ratio (PSNR) of 46 dB, for the OOD validation set of 46 dB and for the OOD test set of 40 dB. These results demonstrate the model's capability to successfully encode image information while retaining a holistic understanding in order to generalize to unseen domains. Fig. 3 illustrates this qualitatively for 5 distinct samples from each hospital and dataset split.
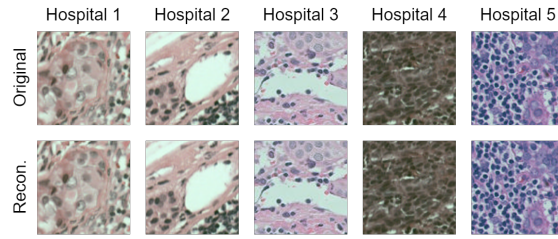
Fig. 3: Qualitative evaluation of our method's reconstruction capability on the CAMELYON17-WILDS dataset.

We also assess the image quality of synthetic images, which exhibit the same anatomy but varied characteristics, generated by our image synthesizer. Fig. 4 demonstrates this process, utilizing randomly extracted patch characteristics for each row. Although our method's patch-wise image reconstruction may produce slight grid artifacts, the synthetic images accurately preserve the original anatomy while displaying uniformly the applied characteristics from the extracted patch. This approach facilitates the generation of a diverse array of samples by altering colorization while maintaining diagnostically relevant anatomy.
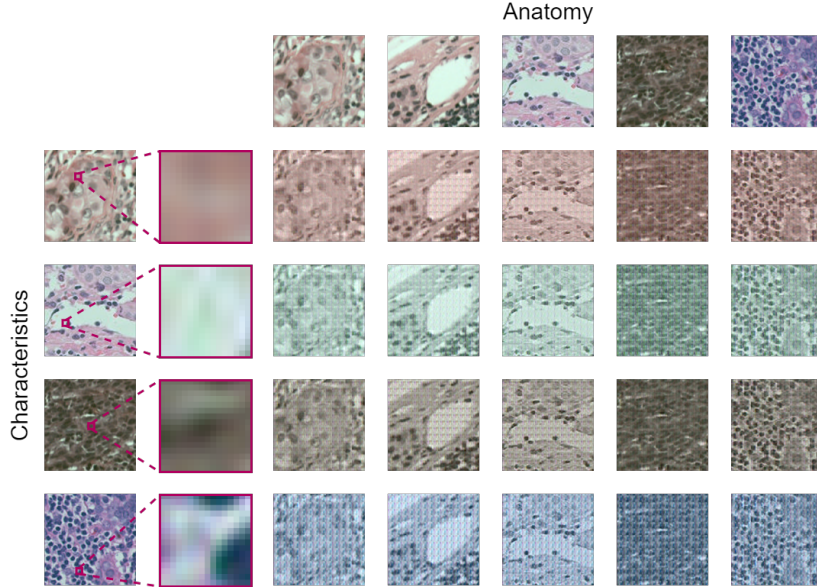

Fig. 4: Qualitative evaluation of the method's generative capabilities on the CAMELYON17-WILDS dataset by means of synthetic images created through its anatomy-characteristics intermixing.

3.2 Disease Classification

To evaluate our method's suitability for improving domain generalization, we employ our stand-alone encoder to generate additional synthetic images with mixed anatomy and characteristics, augmenting the training set diversity on the fly. These synthetic images, alongside the originals, are afterward fed into a subsequent classifier allowing it to learn from a more diverse set of samples, thereby generalizing better to unseen images. For the classifier, we use the same DenseNet-121 architecture [7] used by the baseline methods in WILDS [9]. We evaluate our method on the class-balanced CAMELYON17-WILDS validation and test sets against the top-performing methods from the WILDS leaderboard[1], which utilize the same classifier. The results shown in Table 1 reveal our method's superior accuracy on both sets, setting a new state-of-the-art standard.

Table 1: Accuracy in % on the validation and test set of CAMELYON17-WILDS.

Methods	Val (OOD)	Test (OOD)
ERM[9]	85.80	70.80
LISA[25]	81.80	77.10
ERM with targeted augmentation[5]	92.70	92.10
MBDG[19]	88.10	93.30
ContriMix[17]	91.90	94.60
Ours	**94.16**	**95.44**

We further evaluate our method for the binary classification task of the adapted epithelium-stroma dataset. For this, we train it once on NKI and evaluate it for VGH (val) and IHC (test), as well as train it on VGH and evaluate it for NKI (val) and IHC (test), respectively. We compare the performance against the three domain adaptation methods referenced in [8]. The consistent performance of our method across these evaluations, as presented in Table 2, confirms its strong generalizability potential, clearly outperforming the state of the art.

Table 2: Accuracy in % on the epithelium-stroma dataset.

Methods	Training NKI		Training VGH	
	VGH	IHC	NKI	IHC
DLID[3]	75.70	56.39	86.70	57.36
DDA[6]	77.50	73.17	81.00	52.46
CKA[8]	77.75	73.19	80.17	59.44
Ours	**93.72**	**85.39**	**88.47**	**86.12**

---

[1] https://wilds.stanford.edu/leaderboard/#camelyon17

3.3   Scalability Potential

Finally, we investigate the scalability potential of our method to enhance its reconstruction and image synthesis capabilities. First, we exploit the label-free nature of our encoder (E), enabling the inclusion of unlabeled samples alongside labeled ones during training. This approach allows E to learn from a larger more diverse dataset. To evaluate this, we augment our training data with an additional 302,436 (same amount as labeled training samples) randomly selected samples from the 1,799,247 unlabeled samples available in the Camelyon17-wilds dataset [20]. Through this augmentation, our encoder achieves improved reconstruction performance compared to the base model: 49 dB versus 46 dB for the training set, 49 dB versus 46 dB for the validation set, and 44 dB versus 40 dB for the test set. Furthermore, leveraging a Vision Transformer (ViT) backbone allows us to readily increase model capacity by replacing the ViT-B/16 backbone (86M parameters) with the deeper and more sophisticated ViT-L/16 (322M parameters). Notably, we extend the embedding dimension from 768 to 1,056 to accommodate the requirements of our image synthesizer's matrix multiplication. Training the adapted ViT-L/16 backbone for 10 epochs on Camelyon17-wilds already yields enhanced results, with a reconstruction performance of 49 dB versus 46 dB for the training set, 49 dB versus 46 dB for the validation set, and 42 dB versus 40 dB for the test set. These findings demonstrate that both scaling approaches result in superior performance compared to the base method, underscoring the method's scalability potential in terms of utilizing unlabeled samples and adopting more sophisticated network architectures.

4   Discussion and Conclusion

In this work, we introduce a novel self-supervised, generative method for domain generalization. By employing the power of a Vision Transformer encoder, we successfully generate synthetic images featuring diverse combinations of anatomy and image characteristics in a self-supervised fashion. This approach enriches the representativeness of the dataset without necessitating any domain-specific information, thereby enabling more effective adaptation to previously unseen domains. Through quantitative experimentation on two distinct histopathology datasets, we demonstrate the efficacy of our method. Our qualitative assessment emphasizes the model's proficiency in encoding image data and its capacity to generalize across domains. Moreover, the synthetic images generated by our method faithfully preserve original anatomical details while augmenting dataset diversity. Furthermore, by enabling the utilization of unlabeled samples or the adoption of more sophisticated ViT backbone architectures, our method demonstrates scalability potential, exhibiting improved reconstruction performance and adaptability. We believe that our method's flexibility should allow its application across various modalities for addressing generalization challenges not only in histopathology but also in other applications.

References

1. Beck, A.H., Sangoi, A.R., Leung, S., Marinelli, R.J., Nielsen, T.O., van de Vijver, M.J., West, R.B., van de Rijn, M., Koller, D.: Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Science Translational Medicine **3**(108), 108ra113–108ra113 (2011)
2. Chang, J.R., Wu, M.S., Yu, W.H., Chen, C.C., Yang, C.K., Lin, Y.Y., Yeh, C.Y.: Stain mix-up: Unsupervised domain generalization for histopathology images. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12903 LNCS**, 117–126 (2021)
3. Chopra, S., Balakrishnan, S., Gopalan, R.: Dlid: Deep learning for domain adaptation by interpolating between domains. In: International Conference on Machine Learning Workshop Representation Learning (2013)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
5. Gao, I., Sagawa, S., Koh, P.W., Hashimoto, T., Liang, P.: Out-of-distribution robustness via targeted augmentations. In: NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications (2022)
6. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: International Conference on Machine Learning (2011)
7. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2261–2269 (2016)
8. Huang, Y., Zheng, H., Liu, C., Ding, X., Rohde, G.K.: Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. IEEE Journal of Biomedical and Health Informatics **21**(6), 1625–1632 (2017)
9. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S.M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., Liang, P.: Wilds: A benchmark of in-the-wild distribution shifts. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 5637–5664 (2021)
10. Lafarge, M.W., Pluim, J.P., Eppenhof, K.A., Moeskops, P., Veta, M.: Domain-adversarial neural networks to address the appearance variability of histopathology images. Lecture Notes in Computer Science **10553 LNCS**, 83–91 (2017)
11. Linder, N., Konsti, J., Turkki, R., Rahtu, E., Lundin, M., Nordling, S., Haglund, C., Ahonen, T., Pietikäinen, M., Lundin, J.: Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagnostic Pathology **7**, 1–11 (2012)
12. Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., Hipp, J.D., Peng, L.H., Stumpe, M.C.: Detecting cancer metastases on gigapixel pathology images. ArXiv **abs/1703.02442** (2017)
13. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with restarts. ArXiv **abs/1608.03983** (2016)

14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017)
15. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 1107–1110 (2009)
16. Moscalu, M., Moscalu, R., Dascălu, C.G., Țarcă, V., Cojocaru, E., Costin, I.M., Țarcă, E., Șerban, I.L.: Histopathological images analysis and predictive modeling implemented in digital pathology—current affairs and perspectives. Diagnostics **13** (2023)
17. Nguyen, T.H., Juyal, D., Li, J., Prakash, A., Nofallah, S., Shah, C., Gullapally, S.C., Yu, L., Griffin, M., Sampat, A., Abel, J., Lee, J., Taylor-Weiner, A.: Contrimix: Unsupervised disentanglement of content and attribute for domain generalization in microscopy image analysis (2023)
18. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications **21**(5), 34–41 (2001)
19. Robey, A., Pappas, G.J., Hassani, H.: Model-based domain generalization. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021)
20. Sagawa, S., Koh, P.W., Lee, T., Gao, I., Xie, S.M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., Beery, S., David, E., Stavness, I., Guo, W., Leskovec, J., Saenko, K., Hashimoto, T., Levine, S., Finn, C., Liang, P.: Extending the WILDS benchmark for unsupervised adaptation. In: International Conference on Learning Representations (2022)
21. Stacke, K., Eilertsen, G., Unger, J., Lundstrom, C.: Measuring domain shift for deep learning in histopathology. IEEE Journal of Biomedical and Health Informatics **25**, 325–336 (2021)
22. Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., Litjens, G., van der Laak, J., Ciompi, F.: Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. IEEE Transactions on Medical Imaging **37**(9), 2126–2136 (2018)
23. Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE Transactions on Medical Imaging **35**, 1962–1971 (2016)
24. Wang, Y., Liu, L., Wang, C.: Trends in using deep learning algorithms in biomedical prediction systems. Frontiers in Neuroscience **17** (2023)
25. Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., Finn, C.: Improving out-of-distribution robustness via selective augmentation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 25407–25437. PMLR (2022)