STYLIZING VIT: ANATOMY-PRESERVING INSTANCE STYLE TRANSFER FOR DOMAIN GENERALIZATION

Sebastian Doerrich *Francesco Di Salvo* *Jonas Alle* *Christian Ledig*

xAILab Bamberg, University of Bamberg, Bamberg, Germany

ABSTRACT

Deep learning models in medical image analysis often struggle with generalizability across domains and demographic groups due to data heterogeneity and scarcity. Traditional augmentation improves robustness, but fails under substantial domain shifts. Recent advances in stylistic augmentation enhance domain generalization by varying image styles but fall short in terms of style diversity or by introducing artifacts into the generated images. To address these limitations, we propose *Stylizing ViT*, a novel Vision Transformer encoder that utilizes weight-shared attention blocks for both self- and cross-attention. This design allows the same attention block to maintain anatomical consistency through self-attention while performing style transfer via cross-attention. We assess the effectiveness of our method for domain generalization by employing it for data augmentation on three distinct image classification tasks in the context of histopathology and dermatology. Results demonstrate an improved robustness (up to +13 % accuracy) over the state of the art while generating perceptually convincing images without artifacts. Additionally, we show that *Stylizing ViT* is effective beyond training, achieving a 17 % performance improvement during inference when used for test-time augmentation. The source code is available at https://github.com/sdoerrich97/stylizing-vit .

***Index Terms*—** style transfer, domain generalization, cross-attention, data augmentation, vision transformer

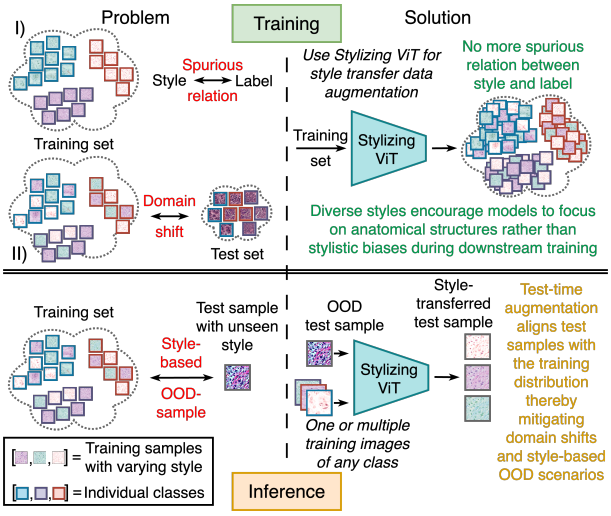

Fig. 1. Overview of how *Stylizing ViT* is used to improve domain generalization during training and inference. During training, it generates stylistically diverse but anatomically consistent images, encouraging downstream classifiers to learn structure-aware representations. At test time, it is used to align unseen input styles with the training distribution, thereby mitigating style-induced domain shifts.

1. INTRODUCTION

Deep learning has achieved remarkable success in image analysis, text processing, and even computer assisted interventions, yet its integration into clinical practice remains limited [1]. A major challenge in medical applications stems from the diverse, heterogeneous, and inherently scarce nature of medical data [2, 3], which affects the generalizability of deep learning models across institutions, domains, and demographic groups [4]. Traditional data augmentation techniques, particularly those utilizing modality-specific characteristics like color gradients or artifacts, improve model robustness by diversifying training data but struggle with substantial domain shifts [5, 6, 7]. In response, recent work suggests that stylistic data augmentation enhances domain generalization by altering image styles [8, 9]. However, these methods either offer limited style diversity or introduce artifacts into the generated images.

In this work, we address these challenges by building on advances in neural style transfer. Neural style transfer aims to render a natural image in the appearance of an artwork while preserving its underlying structure. The seminal work of Gatys et al. [10] introduced this concept using Gram-matrix–based optimization, but its iterative nature made it computationally expensive and slow. This limitation motivated the development of real-time, arbitrary style transfer methods such as AdaIN [11] and AdaAttN [12] where any style can be transferred without having to retrain the model. While early approaches relied on convolutional neural networks (CNNs), recent works increasingly adopt Transformers for their reduced content bias and superior capacity to model long-range dependencies [13]. However, most methods still require a combination of two encoders and a decoder, resulting in a considerable computational overhead.

To tackle these limitations, we introduce *Stylizing ViT*, a novel style transfer method based on a single-encoder Vision Transformer (ViT). Our architecture employs weight sharing within a unified attention block to jointly fuse self- and cross-attention, enabling effective style transfer while preserving anatomical fidelity. We show that *Stylizing ViT* generates realistic, anatomically accurate images with diverse styles across dataset splits and domains. This makes it an effective augmentation technique for augmenting training images on-the-fly, thereby encouraging models to focus on anatomical structures rather than stylistic biases during downstream training. We evaluate this aspect of *Stylizing ViT* across two histopathology datasets and a dermatology one, showing improved performance across domains including varying staining techniques and skin tones. Furthermore, we demonstrate that our approach is able to generalize to unseen anatomical and stylistic variations, highlighting its applicability beyond training-time augmentation, e.g., for inference. Figure 1 illustrates these two key application scenarios side-by-side.

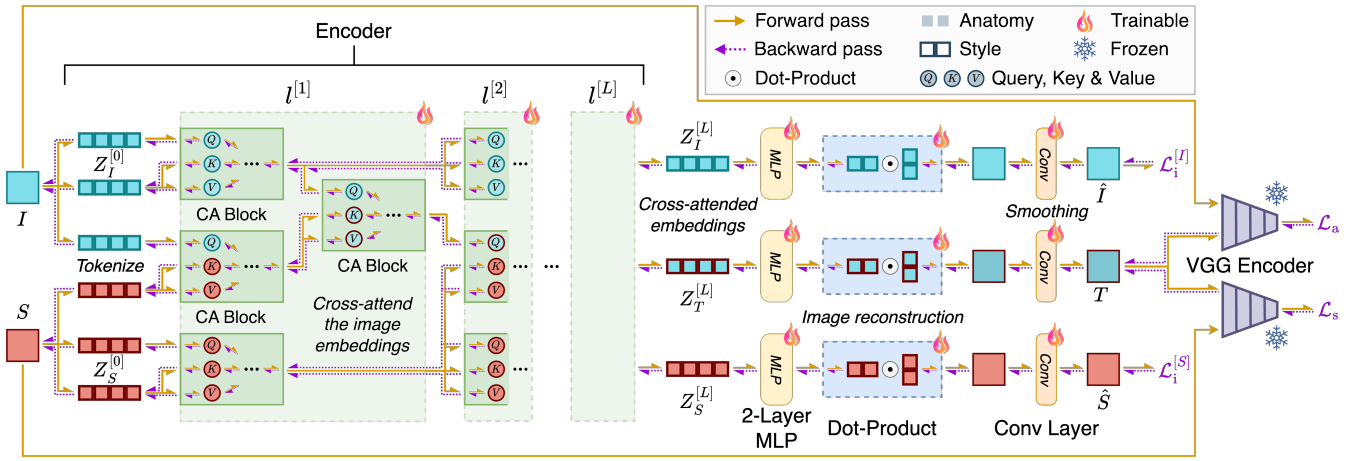

Fig. 2. Overview of our proposed *Stylizing ViT*. The input image pair (I, S) is processed by the encoder that fuses the anatomical structure of I with the style characteristics of S using cross-attention. Subsequently the stylized image T is reconstructed through a two-layer MLP, a dot-product operation, and a convolutional layer. A frozen VGG19 encoder is used during training to compute perceptual losses.

Our main contributions include:

- A modality-agnostic style augmentation method for domain generalization.
- A novel ViT design enabling weight sharing within a unified attention block for both self- and cross-attention information fusion, termed *Stylizing ViT*.
- Applicability for test-time augmentation (TTA) to enhance generalization to new, unseen domains during inference.
- Comprehensive experiments across three medical imaging datasets, demonstrating consistent improvements in style transfer quality, state-of-the-art classification performance (up to +9 % accuracy over prior best), and significant gains from TTA.

2. METHOD

The overview of our method is illustrated in Figure 2. Given two input images, I and S, the objective is to transfer the style of S onto I while preserving I's anatomical integrity. This is achieved through a sequence of model components, including a ViT encoder, a two-layer MLP, a dot-product operation on the resulting feature embeddings, and a final convolutional layer. Specifically, I and S are first processed by a single, shared ViT encoder, where all self-attention operations are replaced with cross-attention to enable the alignment of I's anatomical structure with the style of S. The subsequent MLP scales the feature embeddings for the image reconstruction via the dot-product which restores the original image dimensions. Finally, a convolutional layer is applied to learn more consistent outputs. By combining cross-attention for stylization with robust loss functions from a frozen VGG19 encoder, our method achieves high-quality stylization results without requiring a decoder for reconstruction.

2.1. Stylizing ViT for Style Transfer

We utilize the structural similarity between cross-attention (CA) and self-attention (SA) to introduce *Stylizing ViT*, a Vision Transformer [14] variant that enables the stylization of an image I with the style of a reference image S to produce the style transferred image T. Specifically, we modify the standard ViT encoder by replacing all SA operations with CA, allowing for controlled fusion of anatomical and stylistic information. Furthermore, this allows to employ the same, weight-shared attention block either in the mode of SA or CA. Thus, a single block first processes each input image separately via its SA mode to preserve distinct structural and stylistic representations for the next layer while simultaneously the same block is used to blend the anatomical structure of I with the style attributes of S through CA. Specifically, for tokenized embeddings Z_I and Z_S of I and S, each encoder layer l applies the same, weight-shared CA block in three stages: (i) CA(Z_I, Z_I) which is equivalent to SA(Z_I), (ii) CA(Z_S, Z_S) which is equivalent to SA(Z_S), and (iii) CA(Z_I, Z_S) to produce the stylized output Z_T. A final CA step, using still the same, weight-shared block, further aligns Z_I and Z_T for $l < L$ to ensure anatomical preservation during the stylization.

2.2. Image Reconstruction

To reconstruct the original images I and S as well as the stylized image T, the output embeddings Z_I, Z_S, and Z_T from the encoder are first processed by a shared two-layer MLP. Each embedding is then split into two halves ($z_\mathrm{A}, z_\mathrm{B} \in \mathbb{R}^{n\times d/2}$) along embedding dimension d, shaped into matrix form ($Z_\mathrm{A} \in \mathbb{R}^{n\times c\times p\times v}$, $Z_\mathrm{B} \in \mathbb{R}^{n\times c\times v\times p}$), and multiplied via dot product multiplication along v. Here, n represents the number of image patches, c the channel dimension, p the patch size, and $v = d/(c\times p)$ the hidden dimension for the dot product. To restore the original image dimension the initial patchification from the ViT encoder is reversed. The final convolutional layer with a kernel size of five is used to learn more consistent reconstructions.

2.3. Network Optimization

The stylized image T should preserve the anatomical structures of I while incorporating the style patterns of S. To achieve this, we employ a strategy common in neural style transfer methods [11, 12, 13]. Specifically, we define two perceptual loss terms: an anatomy loss $\mathcal{L}_\mathrm{a}$ that quantifies structural differences between T and I, and a style loss $\mathcal{L}_\mathrm{s}$ that measures stylistic discrepancies between T and S. Additionally, we incorporate the consistency loss $\mathcal{L}_\mathrm{c}$ and identity loss $\mathcal{L}_\mathrm{i}$ from [13] to guide the network in learning accurate representations. All loss terms, except for $\mathcal{L}_\mathrm{i}$, are computed using intermediate feature maps $k \in K$ extracted from a frozen, ImageNet-pretrained VGG19 model, aligning with [10]. The individual loss functions

Table 1. Quantitative evaluation of reconstruction and style transfer quality on the full training sets of Camelyon17-WILDS, Epithelium-Stroma, and Fitzpatrick17k. Reconstruction quality is assessed using PSNR and SSIM on identical image pairs while style transfer quality is evaluated using FID, LPIPS, and ArtFID on distinct image pairs. Higher values (↑) indicate better performance for PSNR and SSIM, while lower values (↓) are better for FID, LPIPS, and ArtFID. The **best performance** for each metric is highlighted in bold.

Method	Camelyon17-WILDS					Epithelium-Stroma					Fitzpatrick17k				
	Reconstruct ↑		Style Transfer ↓			Reconstruct ↑		Style Transfer ↓			Reconstruct ↑		Style Transfer ↓		
	PSNR	SSIM	FID	LPIPS	ArtFID	PSNR	SSIM	FID	LPIPS	ArtFID	PSNR	SSIM	FID	LPIPS	ArtFID
AdaIN	23.7	0.75	55.9	0.17	66.6	23.6	0.85	22.9	0.13	26.9	20.4	0.61	43.5	0.21	54.0
IEContrAST	27.4	0.82	138.7	0.33	185.4	29.3	0.91	36.4	0.22	45.4	30.4	0.77	60.9	0.39	85.9
SANet	26.8	0.81	113.8	0.29	148.6	30.4	0.91	35.1	0.21	43.8	29.4	0.76	81.4	0.44	118.7
StyTR2	30.0	0.88	84.1	0.29	109.4	30.8	0.92	28.4	0.20	35.2	29.4	0.75	104.5	0.47	154.8
SGViT	**46.7**	**0.99**	172.4	0.34	232.7	**47.6**	**0.99**	175.6	0.49	262.92	**40.5**	**0.97**	226.0	0.54	348.8
Stylizing ViT	45.4	**0.99**	**6.2**	**0.06**	**7.6**	39.0	**0.99**	**1.5**	**0.02**	**2.6**	31.5	0.77	**36.9**	**0.20**	**45.5**

$\mathcal{L}_\text{i}, \mathcal{L}_\text{c}, \mathcal{L}_\text{a}, \mathcal{L}_\text{s}$ as well as the total loss $\mathcal{L}_\text{t}$ can be written as:

$$\mathcal{L}_\text{i} = \|I - \hat{I}\|_2 + \|S - \hat{S}\|_2$$

$$\mathcal{L}_\text{c} = \sum \|\phi_k(I) - \phi_k(\hat{I})\|_2 + \|\phi_k(S) - \phi_k(\hat{S})\|_2$$

$$\mathcal{L}_\text{a} = \sum \|\phi_k(I) - \phi_k(T)\|_2$$

$$\mathcal{L}_\text{s} = \sum \|\mu(\phi_k(S)) - \mu(\phi_k(T))\|_2 + \|\sigma(\phi_k(S)) - \sigma(\phi_k(T))\|_2$$

$$\mathcal{L}_\text{t} = \lambda_\text{i}\mathcal{L}_\text{i} + \lambda_\text{c}\mathcal{L}_\text{c} + \lambda_\text{a}\mathcal{L}_\text{a} + \lambda_\text{s}\mathcal{L}_\text{s}$$

where $\hat{I}$ and $\hat{S}$ represent the reconstructions of I and S, respectively, $\phi_k(\cdot)$ the feature embeddings from the k-th layer, and $\mu(\cdot)$ and $\sigma(\cdot)$ the mean and standard deviation of the extracted features. The weights $\lambda_\text{i}, \lambda_\text{c}, \lambda_\text{a}, \lambda_\text{s}$ are set to 70, 1, 7, 10, in alignment with [13].

For our *Stylizing ViT*, we adopt a ViT-B/16 backbone and train it independently on each dataset's training split for 50 epochs with a batch size of 64 using AdamW (learning rate of 0.001) and a cosine annealing schedule. To obtain distinct anatomy-style pairs, we employ a self-supervised approach: the same image batch is loaded twice, with the first batch being used for the anatomy images while the samples in the second batch, the style images, are shuffled to avoid corresponding anatomy-style pairs of identical samples.

3. EXPERIMENTS AND RESULTS

We evaluate our method for (1) its effectiveness in transferring styles while preserving anatomical structures, (2) its utility as a data augmentation technique for enhancing domain generalization, and (3) its potential for TTA. Our evaluation employs two histopathology and one dermatology dataset. The first, Camelyon17-WILDS [15], comprises H&E-stained lymph node image patches for tumor identification across five hospitals (H1-H5). We adopt the official dataset splits: train (H1-H3; 302,436 samples), val (H4; 34,904), and test (H5; 85,054). The second histopathology dataset, sourced from [16], aggregates three public datasets for epithelium-stroma classification, namely NKI [17] (train: 8,337), VGH [17] (val: 5,920), and IHC [18] (test: 1,376). NKI and VGH contain H&E-stained breast cancer images, while IHC consists of IHC-stained colorectal cancer images. Finally, we utilize Fitzpatrick17k [19], a dataset of dermatology images labeled with Fitzpatrick skin tones (I-VI). Following [20], we define three groups: {I-II: train (7,755), III-IV: val (6,089), V-VI: test (2,168)}, representing progressively darker skin tones. All datasets are standardized to a 224×224 resolution using bicubic interpolation while maintaining aspect ratios.

3.1. Training Set Stylization

To evaluate the effectiveness of our method in transferring style while preserving anatomical fidelity, we assess both reconstruction and style transfer quality on the full training sets. For reconstruction, we compute the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) which quantify the preservation of fine-grained structural detail on pairs of identical images. For style transfer evaluation, we use pairs of distinct images and compute the Fréchet Inception Distance (FID) [21], Learned Perceptual Image Patch Similarity (LPIPS) [22], and ArtFID [23]. FID measures how closely the stylized image matches the style distribution of the style image, LPIPS quantifies content fidelity between the stylized and content image, and ArtFID jointly captures both content preservation and stylistic accuracy, aligning well with human judgment [24].

We compare our method against four state-of-the-art neural style transfer methods: AdaIN [11], IEContrAST [25], SANet [26], and StyTR^2 [13] as well as the recent SGViT [8]. All of these were trained per dataset using their official implementations and recommended training settings.

Quantitative results are presented in Table 1, with representative qualitative examples shown in Figure 3. Our method consistently outperforms all reference methods in style transfer metrics (FID, LPIPS, ArtFID) across all datasets while achieving the best or second best reconstruction scores (PSNR, SSIM). This indicates that our approach successfully applies diverse styles to training samples without compromising anatomical structure, an essential requirement for medical image augmentation. In contrast, the reference methods either fail to reliably transfer style or introduce severe anatomical distortions.

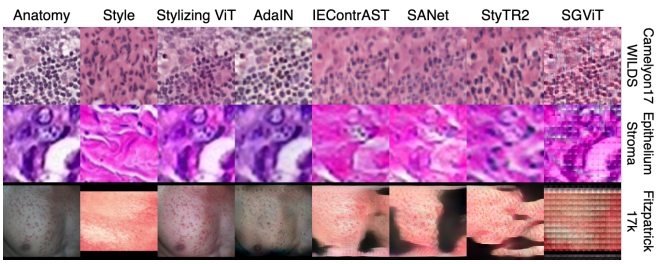

Fig. 3. Qualitative comparison of style transfer quality on training image pairs. Our method consistently outperforms all reference methods in applying diverse styles without compromising anatomy.

Table 2. Test accuracy (%) of a DenseNet121 disease classifier averaged over three runs per dataset. The Baseline corresponds to training with the optimal dataset-specific conventional augmentation strategy. Each subsequent row shows the result when additionally employing the respective style transfer method. The **best-performing method** per dataset is highlighted in bold.

Method	Camelyon17 WILDS	Epithelium Stroma	Fitzpatrick 17k
Baseline	90.24±2.4	76.09± 3.3	81.18±1.1
+ AdaIN	93.36±1.1	66.23±20.8	**81.52±0.3**
+ IEContrAST	95.24±1.1	63.40±20.2	80.58±0.4
+ SANet	95.62±0.3	58.77±13.5	81.33±0.4
+ StyTR2	93.86±0.4	69.67± 3.3	80.89±0.3
+ SGViT	95.25±0.5	58.19± 9.7	80.40±0.6
+ Stylizing ViT	**95.65±0.9**	**89.07± 0.7**	80.92±0.6

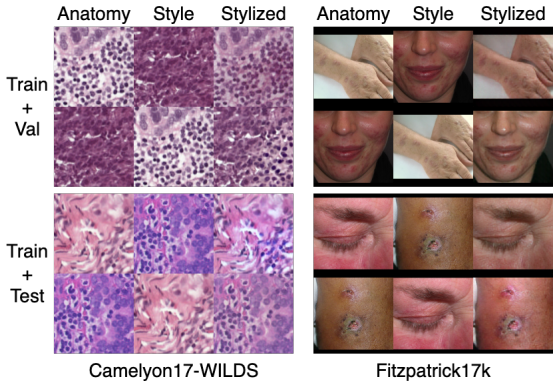

Fig. 4. Style transfer results for Camelyon17-WILDS (left) and Fitzpatrick17k (right), showing high quality images across different dataset splits and domains such as staining variations and skin tones.

3.2. Disease Classification

We evaluate the effectiveness of our method for domain generalization and its compatibility with traditional augmentation strategies by integrating it into the training pipeline of a DenseNet121 disease classifier. The core intuition is that *Stylizing ViT* should complement rather than replace established augmentation strategies. These conventional techniques address common variations in scale, color, and illumination, while our approach enhances domain generalization by generating anatomically consistent yet stylistically diverse samples. This encourages the model to prioritize structural representations over domain-specific appearance cues, a property particularly beneficial in histopathology and dermatology, where staining differences (Camelyon17-WILDS, Epithelium-Stroma) and skin tone variations (Fitzpatrick17k) often impede cross-domain robustness.

Following the WILDS protocol [15], we report test accuracy at a fixed threshold of 0.5. We compare *Stylizing ViT* against all previously introduced style transfer methods across three datasets, each representing a distinct generalization challenge. For each dataset, we first determine a baseline based on the optimal conventional augmentation strategy. We consider resized cropping, grayscale conversion, color jitter, RandAugment, and domain-specific augmentations from [27]. We identified color jitter (Camelyon17-WILDS, Fitzpatrick17k) and resized cropping (Epithelium-Stroma) as optimal methods for the respective datasets. We then quantify the benefit of each style transfer method, including ours, by employing them in addition to the selected optimal dataset-specific augmentation.

The DenseNet121 classifier is trained for 100 epochs and three independent runs under each augmentation configuration using AdamW with cosine annealing (initial learning rate 0.001), batch size 256, and early stopping. Dataset-specific augmentations are applied online with default parameters, while style transfer is applied with probability 0.33 to balance exposure to stylized and original samples. All methods perform full image-to-image transfer except SGViT, which transfers a single style patch across all anatomical regions, emphasizing local texture adaptation and currently representing the state of the art on Camelyon17-WILDS. Table 2 summarizes the results. *Stylizing ViT* achieves state-of-the-art performance on Camelyon17-WILDS and yields a substantial (+13 %) improvement over the next-best method on Epithelium-Stroma. On Fitzpatrick17k, however, none of the evaluated methods produce consistent gains, indicating that this dataset presents challenges that training-time augmentation alone may not adequately address.

3.3. Test-Time Augmentation

To evaluate the generalization capabilities of our method beyond the anatomies and styles encountered during training, we explore its application for augmenting validation and test images, potentially containing styles never seen by the encoder during training. Specifically, we interchange anatomy and style images across all three dataset splits. Figure 4 presents stylized results for Camelyon17-WILDS and Fitzpatrick17k, demonstrating high-quality style transfer across all splits. These findings highlight the potential of our method for test-time augmentation, where test images are augmented with training styles to better align them with the training distribution. We quantitatively evaluate the applicability of our method for this on Camelyon17-WILDS by stylizing each test image with the style of a randomly selected training image. The stylized images are then classified using the DenseNet121 model from before trained without any data augmentation. Our results indicate a marginally significant improvement (paired t-test, $p < 0.1$) despite the small sample size (3 runs) and large variance in the baseline performance, increasing accuracy from $59.64\,\%$ to $77.40\,\%$. These findings emphasize the effectiveness of our approach to enhance generalization at test time.

4. DISCUSSION AND CONCLUSION

We introduced *Stylizing ViT*, a novel Vision Transformer architecture that unifies self- and cross-attention within a single block, enabling anatomically consistent style transfer for visual domains. Our experiments show that this mechanism can be effectively utilized for data augmentation, improving robustness to domain shifts. Furthermore, *Stylizing ViT* enhances domain generalization at test time by aligning unseen inputs more closely with the training distribution.

While training can be computationally demanding, primarily due to the joint attention mechanism, our approach remains practical in real-world settings. Our stylization operates within milliseconds per batch and can be seamlessly integrated into the data loading pipeline, making it well suited for efficient, on-the-fly augmentation during training. In terms of performance, *Stylizing ViT* achieves state-of-the-art results on Camelyon17-WILDS and substantial improvements on Epithelium-Stroma. Although its impact is more limited on Fitzpatrick17k, qualitative results remain convincing. This suggests that datasets involving complex, subtle, or patient-specific domain shifts, such as variations in skin tone, lighting, or photographic conditions, may require additional strategies.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [15], [17], [18], and [19]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

This work was funded through the Hightech Agenda Bayern (HTA) of the Free State of Bavaria, Germany.

7. REFERENCES

[1] Karin Stacke, Gabriel Eilertsen, Jonas Unger, et al., "Measuring domain shift for deep learning in histopathology," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 325–336, 2021.

[2] Amjad Khan, Andrew Janowczyk, Felix Müller, et al., "Impact of scanner variability on lymph node segmentation in computational pathology," *Journal of Pathology Informatics*, 2022.

[3] Raphael Schäfer, Till Nicke, Henning Höfener, et al., "Overcoming data scarcity in biomedical imaging with a foundational multi-task model," *Nature Computational Science*, vol. 4, no. 7, pp. 495–509, 2024.

[4] Aidan Dulaney and John Virostko, "Disparities in the demographic composition of the cancer imaging archive," *Radiology: Imaging Cancer*, vol. 6, no. 1, pp. e230100, 2024.

[5] Tauhidul Islam, Md. Sadman Hafiz, Jamin Rahman Jim, et al., "A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions," *Healthcare Analytics*, vol. 5, pp. 100340, 2024.

[6] Cheng Ouyang, Chen Chen, Surui Li, et al., "Causality-inspired single-source domain generalization for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 1095–1106, 2023.

[7] Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, et al., "Domain generalization for medical image analysis: A review," *Proceedings of the IEEE*, vol. 112, no. 10, pp. 1583–1609, 2024.

[8] Sebastian Doerrich, Francesco Di Salvo, and Christian Ledig, "Self-supervised vision transformer are scalable generative models for domain generalization," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 2024, pp. 644–654.

[9] Tan H. Nguyen, Dinkar Juyal, Jin Li, et al., "Contrimix: Scalable stain color augmentation for domain generalization without domain labels in digital pathology," in *Proceedings of the MICCAI Workshop on Computational Pathology*, 2024, vol. 254, pp. 121–130.

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, "A neural algorithm of artistic style," *ArXiv*, 2015.

[11] Xun Huang and Serge J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1510–1519.

[12] Songhua Liu, Tianwei Lin, Dongliang He, et al., "Adaattn: Revisit attention mechanism in arbitrary neural style transfer," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[13] Yingying Deng, Fan Tang, Weiming Dong, et al., "Stytr^2: Image style transfer with transformers," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[15] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, et al., "Wilds: A benchmark of in-the-wild distribution shifts," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, vol. 139, pp. 5637–5664.

[16] Chenxin Li, Xin Lin, Yijin Mao, et al., "Domain generalization on medical imaging classification using episodic training with task augmentation," *Computers in Biology and Medicine*, vol. 141, pp. 105144, 2022.

[17] Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, et al., "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*, vol. 3, no. 108, 2011.

[18] Nina Linder, Juho Konsti, Riku Turkki, et al., "Identification of tumor epithelium and stroma in tissue microarrays using texture analysis," *Diagnostic Pathology*, vol. 7, pp. 1–11, 2012.

[19] Matthew Groh, Caleb Harris, Luis R. Soenksen, et al., "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1820–1828.

[20] Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, et al., "Disparities in dermatology ai performance on a diverse, curated clinical image set," *Science Advances*, vol. 8, 2021.

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, et al., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*. 2017, vol. 30, Curran Associates, Inc.

[22] Richard Zhang, Phillip Isola, Alexei A. Efros, et al., "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[23] Matthias Wright and Björn Ommer, "Artfid: Quantitative evaluation of neural style transfer," in *Pattern Recognition*, 2022, pp. 560–576.

[24] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo, "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 8795–8805.

[25] Haibo Chen, Lei Zhao, Zhizhong Wang, et al., "Artistic style transfer with internal-external learning and contrastive learning," in *Advances in Neural Information Processing Systems*. 2021, vol. 34, pp. 26561–26573, Curran Associates, Inc.

[26] Dae Young Park and Kwang Hee Lee, "Arbitrary style transfer with style-attentional networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5873–5881.

[27] Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig, "Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions," 2024.

A. OVERVIEW OF SUPPLEMENTARY MATERIAL

The following sections provide additional experimental details, analyses, and qualitative results that complement the main paper. In Section B, we report implementation details, including computational setup, sources of randomness, data augmentation protocols, and reference style transfer methods used throughout our experiments. In Section C, we illustrate the ability of the proposed method to modify visual appearance while maintaining anatomical structure via a controlled experimental setup. In Section D, we provide an extensive ablation study analyzing the contribution of individual loss terms, architectural components, model size, and the proportion of augmented samples used during downstream training. Finally, in Section E, we include additional qualitative style transfer results across all evaluated datasets and style transfer methods.

B. EXPERIMENTAL DETAILS

B.1. Computation

All experiments were conducted on a single NVIDIA RTX 6000 Ada GPU (48 GB RAM) running Ubuntu 24.04.2 LTS with Python 3.12.

B.2. Randomness

Each result reports the average over three independent runs with different random seeds: $S \in \{265017005, 145288287, 206368124\}$

B.3. Data Augmentation

Traditional augmentations (random resized crop, grayscale, and color jitter) were applied using Torchvision's built-in implementations with defaul hyperparameters. For RandAugment and the domain-specific MedMNIST-C augmentations, we used their official implementations with default hyperparameters.

B.4. Style Transfer Reference Methods

All reference style transfer methods were adapted from their official implementations and trained using their default settings and hyperparameters.

C. ANATOMY-PRESERVING STYLE TRANSFER

We assess whether our method can transfer the style of a source image $\tilde{S}$ onto a target image I while preserving the anatomical content of I. Since style lacks a precise definition in images, we construct a controlled evaluation protocol using a predefined augmentation function $f(\cdot)$ that alters style while preserving content. Specifically, we define a color-altering transform $f(X)$, which converts an image X into $\tilde{X}$. This controlled setup allows us to create both the style reference $\tilde{S} = f(S)$ and target image $\tilde{I} = f(I)$ for evaluating the stylization of I in the style of $\tilde{S}$.

Figure 5 presents this evaluation for a single image pair (I, S) from the Epithelium-Stroma dataset. We apply a color jitter augmentation f to generate the style reference $\tilde{S}$ and the expected target $\tilde{I}$. We then apply our method to transfer the style from $\tilde{S}$ to I, resulting in the stylized output image T. T visually aligns with the expected target $\tilde{I}$, indicating that our method accurately transfers the style of $\tilde{S}$ without distorting the anatomy of I.

Additional qualitative examples from both the Epithelium-Stroma and Camelyon17-WILDS datasets are provided in Figure 6.

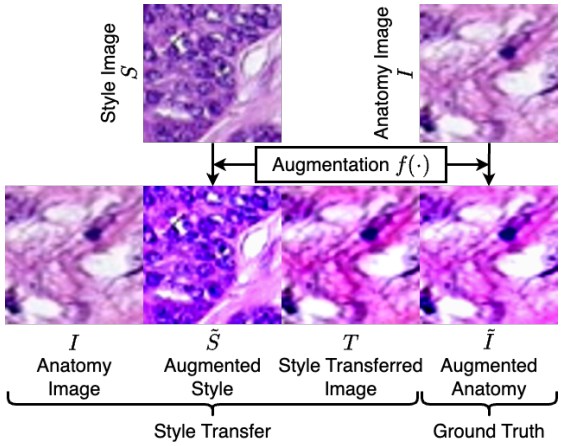

Fig. 5. Demonstration of our method's ability to transfer style while preserving anatomical structure. Given an input image pair (I, S), we first apply a shared color jitter augmentation $f(\cdot)$ to produce $\tilde{I}$ and $\tilde{S}$. Our model generates T, which closely matches $\tilde{I}$, indicating successful transfer of style while maintaining anatomical fidelity.

D. ABLATION STUDY

We conduct an ablation study on the Epithelium-Stroma dataset to quantify the contribution of individual loss terms and model components. For each configuration, we retrain the model from scratch and evaluate both reconstruction quality (using identical input images) and style transfer quality (using distinct anatomy and style images). Reconstruction quality is assessed using PSNR and SSIM, while style transfer quality is evaluated with FID and ArtFID.

In addition, we study the impact of encoder size across all three datasets and analyze how the proportion of stylized samples used during downstream classifier training affects classification performance.

D.1. Loss Terms

Table 3. Impact of individual loss terms on reconstruction and style transfer quality. PSNR and SSIM (↑) measure reconstruction performance while FID and ArtFID (↓) assess style transfer quality.

Method	Reconstruction		Style Transfer	
	PSNR ↑	SSIM ↑	FID ↓	ArtFID ↓
Identity	39.5	0.99	35.7	36.8
+ Consistency	40.8	0.99	35.0	36.1
+ Anatomy	39.5	0.98	33.7	35.0
+ Style	37.7	0.98	28.9	31.4

To analyze the effect of each loss term, we progressively introduce components into the full objective during training. Starting with only the identity loss $\mathcal{L}_{\text{i}}$, we incrementally add the consistency loss $\mathcal{L}_{\text{c}}$, anatomy loss $\mathcal{L}_{\text{a}}$, and finally the style loss $\mathcal{L}_{\text{s}}$.

Reconstruction performance is evaluated on the training set, while style transfer is assessed using the validation set as anatomy input and the training set as style input. This pairing avoids overemphasizing trivial solutions. Quantitative results are reported in Table 3, with representative qualitative examples shown in Figure 7.

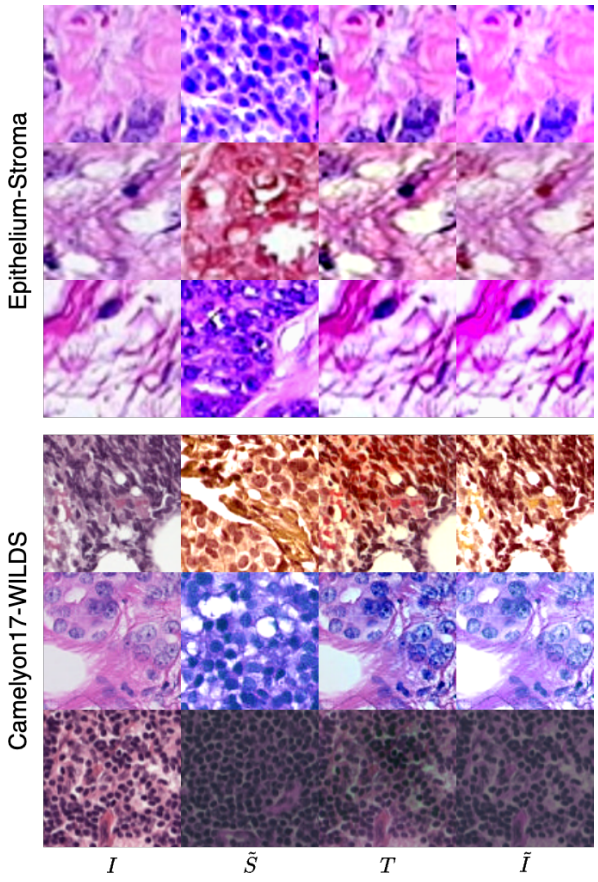

Fig. 6. Additional controlled style transfer examples from the Epithelium-Stroma and Camelyon17-WILDS datasets. For each row, from left to right: (1) anatomy image I, (2) augmented style image $\tilde{S} = f(S)$, (3) stylized output T, and (4) expected target $\tilde{I} = f(I)$. The style transform f corresponds to PyTorch's `ColorJitter` augmentation with default parameters.

The results indicate that the identity and consistency losses primarily drive high-fidelity reconstruction. Incorporating the anatomy loss improves the preservation of fine structural details during style transfer, while the style loss is essential for enabling meaningful stylistic transformation. In its absence, the model tends to revert to reconstructing the anatomy image without effective style modulation.

D.2. Model Components

We further investigate the contribution of each architectural component. Starting from the encoder alone, we incrementally add: (i) the MLP; (ii) the dot-product-based reconstruction; and (iii) the final convolution. Both reconstruction and style transfer performance are evaluated on the training set. Qualitative comparisons are shown in Figure 8, with quantitative results summarized in Table 4.

The results reveal that both the MLP and dot-product are crucial for accurate reconstructions. While the dot-product mechanism slightly degrades style transfer performance, it enhances the preservation of anatomical details. Finally, the inclusion of the convolutional refinement layer leads to improvements in both reconstruction and style transfer quality.

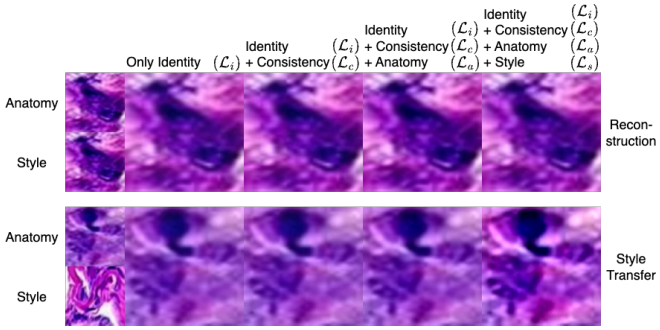

Fig. 7. Qualitative evaluation of reconstruction and style transfer results on Epithelium-Stroma under different loss configurations.

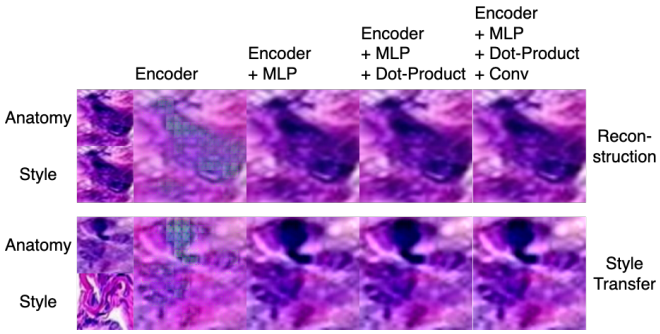

Fig. 8. Qualitative evaluation of reconstruction and style transfer results on Epithelium-Stroma for different model configurations.

D.3. Model Size

We examine the effect of encoder capacity on reconstruction and style transfer performance by replacing the *Stylizing ViT* encoder in its base configuration with two smaller variants: a *tiny* model (3 attention heads, 192-dimensional embeddings) and a *small* model (6 attention heads, 384-dimensional embeddings).

Qualitative and quantitative results across datasets (Figure 9 and Table 5) indicate that the tiny variant consistently underperforms the larger models in both reconstruction and stylization tasks. In contrast, the small variant provides a competitive alternative to the base configuration, particularly on smaller datasets such as Epithelium-Stroma and Fitzpatrick17k. Notably, on Fitzpatrick17k, the small model slightly outperforms the base model in terms of style transfer quality.

Table 4. Impact of individual model components on reconstruction and style transfer quality. PSNR and SSIM (↑) measure reconstruction performance while FID and ArtFID (↓) assess style transfer quality.

Method	Reconstruction		Style Transfer	
	PSNR ↑	SSIM ↑	FID ↓	ArtFID ↓
Encoder	25.7	0.78	49.1	65.9
+ MLP	35.8	0.98	2.1	3.2
+ Dot-Product	36.4	0.98	2.6	3.7
+ Conv-layer	39.0	0.99	1.5	2.6

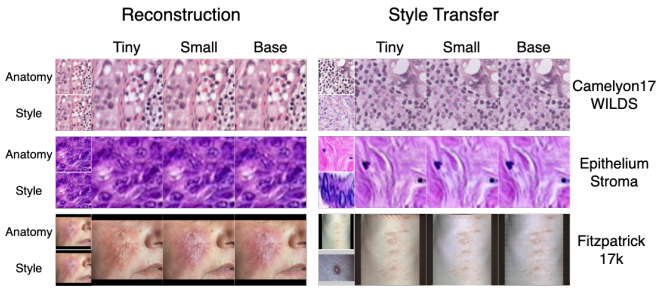

Fig. 9. Qualitative comparison of reconstruction and style transfer results across encoder sizes (tiny, small, base) for exemplary images of Camelyon17-WILDS, Epithelium-Stroma, and Fitzpatrick17k.

On Camelyon17-WILDS, however, the base configuration consistently yields the best overall performance, suggesting that higher model capacity is beneficial when sufficient training data are available. Overall, these results demonstrate the flexibility of the proposed framework, which can be adapted to different dataset sizes and computational constraints. In particular, the small variant offers a favorable trade-off between performance and efficiency, making it suited for deployment in resource-constrained settings.

Table 5. Comparison of reconstruction (PSNR, SSIM) and style transfer (FID, LPIPS, and ArtFID) quality of tiny, small, and base *Stylizing ViT* on the training sets of Camelyon17-WILDS, Epithelium-Stroma, and Fitzpatrick17k.

Backbone	Reconstruction		Style Transfer		
	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓	ArtFID ↓
Camelyon17-WILDS					
Tiny	33.4	0.95	10.6	0.08	12.5
Small	41.5	0.98	6.4	0.06	7.9
Base	45.4	0.99	6.2	0.06	7.6
Epithelium-Stroma					
Tiny	30.7	0.96	19.4	0.04	21.1
Small	38.9	0.99	1.5	0.02	2.6
Base	39.0	0.99	1.5	0.02	2.6
Fitzpatrick17k					
Tiny	30.7	0.84	33.2	0.18	40.5
Small	30.6	0.82	21.6	0.16	26.3
Base	31.5	0.77	36.9	0.20	45.5

D.4. Amount of Augmented Samples

We investigate the effect of varying the proportion of augmented samples used during downstream classifier training on the Epithelium-Stroma dataset. Specifically, we compare different augmentation ratios: 0%, 10%, 33%, 50%, and 100%, applied per mini-batch, using stylized samples generated by our method. We evaluate two configurations: (i) pure stylization via our proposed model, and (ii) stylization combined with a resized crop augmentation applied to the style images prior to generation. The results, visualized in Figure 10, indicate that the downstream test accuracy remains remarkably stable once the proportion of augmented samples reaches approximately 33%. Notably, the 100% augmentation setting achieves competitive and in some cases superior performance compared to more conservative configurations such as 33% or 50%. This suggests that our stylization framework is capable of generating sufficiently diverse and label-preserving samples, making even fully augmented batches viable for training without degrading performance.

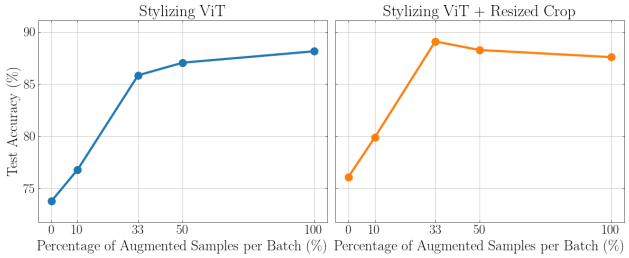

Fig. 10. Impact of varying the proportion of augmented samples per batch on test accuracy for the Epithelium-Stroma dataset.

E. ADDITIONAL STYLE TRANSFER RESULTS

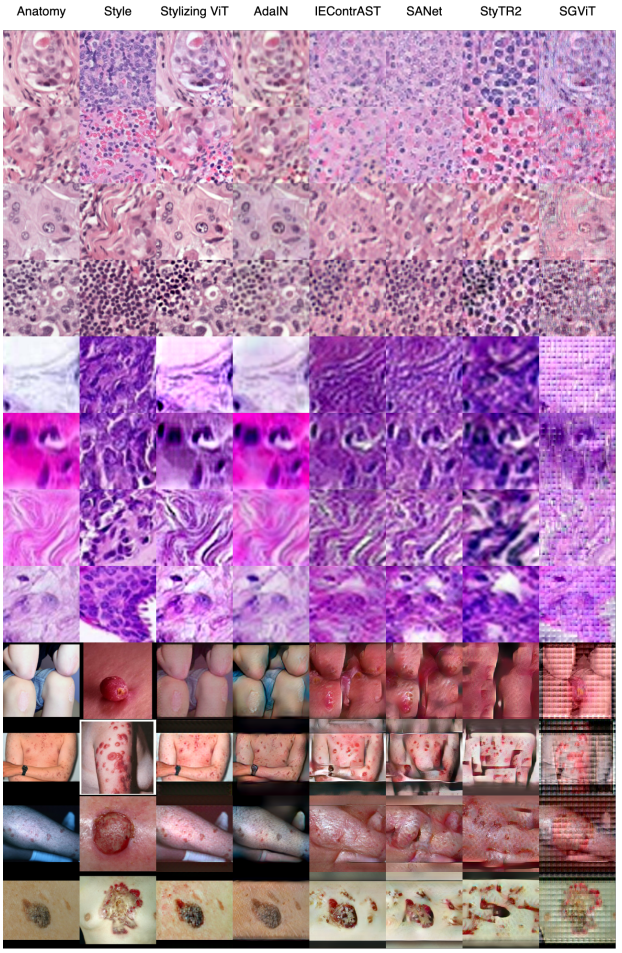

Fig. 11. Additional stylized images generated by all evaluated style transfer methods on training image pairs from Camelyon17-WILDS, Epithelium-Stroma, and Fitzpatrick17k.