Computers in Biology and Medicine 175 (2024) 108410

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Using histopathology latent diffusion models as privacy-preserving dataset augmenters improves downstream classification performance

Jan M. Niehues [a], Gustav Müller-Franzes [b], Yoni Schirris [a,c,d], Sophia Janine Wagner [a,h,i], Michael Jendrusch [e], Matthias Kloor [e], Alexander T. Pearson [f], Hannah Sophie Muti [a,g], Katherine J. Hewitt [a,g], Gregory P. Veldhuizen [a,g], Laura Zigutyte [a], Daniel Truhn [b,1], Jakob Nikolas Kather [a,j,k,l,1,*]

[a] Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany
[b] Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany
[c] Netherlands Cancer Institute, 1066 CX, Amsterdam, the Netherlands
[d] University of Amsterdam, 1012 WP, Amsterdam, the Netherlands
[e] Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany
[f] Department of Medicine, University of Chicago, Chicago, USA
[g] Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany
[h] Helmholtz Munich – German Research Center for Environment and Health, Munich, Germany
[i] School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[j] Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom
[k] Department of Medicine I, University Hospital Dresden, Dresden, Germany
[l] Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

ARTICLE INFO

Keywords:
Artificial intelligence
Generative models
Colorectal cancer
Computational pathology
Diffusion models
Generative adversarial networks

ABSTRACT

Latent diffusion models (LDMs) have emerged as a state-of-the-art image generation method, outperforming previous Generative Adversarial Networks (GANs) in terms of training stability and image quality. In computational pathology, generative models are valuable for data sharing and data augmentation. However, the impact of LDM-generated images on histopathology tasks compared to traditional GANs has not been systematically studied.

We trained three LDMs and a styleGAN2 model on histology tiles from nine colorectal cancer (CRC) tissue classes. The LDMs include 1) a fine-tuned version of stable diffusion v1.4, 2) a Kullback-Leibler (KL)-autoencoder (KLF8-DM), and 3) a vector quantized (VQ)-autoencoder deploying LDM (VQF8-DM). We assessed image quality through expert ratings, dimensional reduction methods, distribution similarity measures, and their impact on training a multiclass tissue classifier. Additionally, we investigated image memorization in the KLF8-DM and styleGAN2 models.

All models provided a high image quality, with the KLF8-DM achieving the best Frechet Inception Distance (FID) and expert rating scores for complex tissue classes. For simpler classes, the VQF8-DM and styleGAN2 models performed better. Image memorization was negligible for both styleGAN2 and KLF8-DM models. Classifiers trained on a mix of KLF8-DM generated and real images achieved a 4% improvement in overall classification accuracy, highlighting the usefulness of these images for dataset augmentation.

Our systematic study of generative methods showed that KLF8-DM produces the highest quality images with negligible image memorization. The higher classifier performance in the generatively augmented dataset suggests that this augmentation technique can be employed to enhance histopathology classifiers for various tasks.

* Corresponding author. Medizinische Fakultät der TU Dresden, EKFZ für Digitale Gesundheit, Postfach 151, Fetscherstraße 74, 01307, Dresden, Germany.
E-mail address: jakob-nikolas.kather@alumni.dkfz.de (J.N. Kather).
[1] shared last authorship.

https://doi.org/10.1016/j.compbiomed.2024.108410
Received 15 October 2023; Received in revised form 23 March 2024; Accepted 2 April 2024
Available online 4 April 2024

1. Introduction

Histopathology is a cornerstone of precision oncology. The diagnosis for virtually every solid tumor is made by a pathologist on histology slides. Histology slides can be digitized and analyzed with Deep Learning (DL) methods, which can extract a wealth of information from histomorphology in, among others, colorectal cancer (CRC) [1–6]. As precision medicine is developing, an increasing number of different biomarkers have become relevant, combined with a global shortage of pathologists. Training the next generation of pathologists is paramount and synthetic images can be utilized in their education [7]. Furthermore, we may automate decision-making by applying DL to digital pathology [8]. However, handling and analyzing digitized pathology slides with associated clinical data is challenging: Both comprise privileged data that cannot be readily shared. Patient privacy as well as legal and ethical quandaries must be considered before use, which constitute barriers to research. Moreover, histopathology image datasets are often imbalanced, with only a few patients being part of a biologically relevant class. In the case of CRC, patients express microsatellite instability (MSI), a highly relevant biomarker, in only 15–20% of cases [9], and DL models may struggle to effectively learn distinctive MSI characteristics. Finally, biomarker prediction from histopathological images is ideally robust, generalizable, and accurate [10]. Classifiers for biomarker prediction should therefore not depend on spurious features in the training data that mitigate performances when deploying models on previously unseen data. In histopathology, some established procedures for data augmentation already exist [11,12], however, the lack of generalization of classifiers remains an issue.

All of the above issues can be addressed with generative deep learning. This technique has been successfully applied to digital pathology: Generative models synthesize pathology images that are realistic [13] and can model complex generative processes like those that translate upstream genomic changes to downstream morphological characteristics without revealing actual patient data [14,15]. At the same time, synthetic images that express characteristic histomorphological features can be used to augment imbalanced datasets. Synthesizing images of a less-represented class (e.g. MSI) has successfully improved DL model performances [14,15]. Finally, generative models have been used to speed up the processing of histopathological slides [16].

So far, most research relies on generative adversarial networks (GAN) [17]. GANs can generate images of high quality and similarity to original images for a large range of different image modalities in cancer [10]. Diffusion models (DM) implemented as denoising diffusion probabilistic models (DDPM) [18,19] have overtaken GANs in terms of training stability and greater image quality with fewer artifacts [20–22]. This approach, however, usually consumes hundreds of GPU days for training, and image generation is time-consuming [23,24]. Recently, latent diffusion models (LDMs) (e.g. stable diffusion [25] and DALLE-2 [26]) have further improved training and generation efficiency [25]. In the medical domain, DMs and LDMs have already been evaluated in radiology [27,28] and the histopathology domain [29]. Furthermore, the superiority in terms of image quality of LDMs compared to GANs has recently been shown for various medical image modalities such as MSI in CRC, eye fundus imaging, and chest X-rays [21]. However, for large DMs trained on small datasets, it has been shown that memorization occurs [30,31]. Therefore, if patient privacy data issues are to be resolved by LDMs, any leakage of original data via memorization should be evaluated and avoided.

A recent work by Ye et al. examined and assessed the possibility of augmenting small-scale datasets with the help of large-scale histopathological pretrained LDMs, and the authors successfully applied their method to downstream classification tasks in CRC [32]. However, this study did neither systematically investigate the effect of different autoencoder architectures, nor examine the potential of using LDMs for sharing of medical image data. In this paper, we perform a systematic evaluation of several recent LDM architectures and compare them to the most commonly used GAN models. We compare four generative models for the synthesis of small 256 × 256 pixel histopathological image patches. The models include three LDMs: two LDMs that use different autoencoder architectures, and a fine-tuned version of stable diffusion v1.4. As a representative of a state-of-the-art GAN, we evaluate the styleGAN2 model, an improved styleGAN architecture [33]. We train all models on the publicly available NCT-CRC-HE-100K dataset [34] that contains histological images from patients with CRC. The analysis is split into three parts. In the first part, we evaluate the generated images qualitatively and quantitatively for each model. For the qualitative evaluation, medical experts are asked to distinguish synthesized images from original images. For the quantitative evaluation, the Frechet Inception Distance (FID) [35] between synthetic and original images is calculated. Additionally, images from the best LDM and styleGAN2 models are visually analyzed in comparison to the original dataset via a t-distributed stochastic neighbor embedding (t-SNE) plot [36]. In the second part, we compare data memorization in the generation of images from the styleGAN2 model with our best LDM. In the last part, we show that augmentation of the original dataset by adding synthetic images from our best LDM improves downstream classification performance on an independent test set. Based on this, we conclude that images from LDMs can be used to augment histopathological training data to improve classification performances and respect patient privacy at the same time.

2. Material and methods

2.1. Datasets

In all our experiments, we train the generative models on the NCT-CRC-HE-100K dataset [34]. The dataset consists of 100,000 hematoxylin & eosin (H&E) stained histopathological tissue tiles of 224 × 224 pixels at 0.5 μm per pixel resolution, from 89 patients with CRC. The tiles are classified into nine different tissue classes: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM). These classes have the following order in complexity as measured by entropy [37,38]: ADI (2.6), BACK (3.2), MUC (5.0), MUS (5.2), STR (5.2), DEB (5.4), LYM (5.4), TUM (5.4), and NORM (5.5). The tiles in the dataset are either Macenko-normalized [11] or show the native H&E color distributions. We train our models on the non-normalized set of tiles. Bilinear interpolation is used to upscale the tiles to a size of 256 × 256 pixels. The separate Macenko-normalized CRC-VAL-HE-7K test set with 7180 tiles extracted from 50 patients is used to evaluate tissue classification performances.

2.2. Generative networks

In this paper, we compare two conceptually different image generation techniques for the synthesis of histopathological images: GANs and LDMs.

2.2.1. styleGAN2

In its simplest form, a GAN consists of two functions that are trained in a competing fashion: The generator (G) generates images from a normally distributed latent space $p_z(\mathbf{z})$ that can be class-conditioned. The discriminator (D) is trained to discriminate synthetic from non-synthetic images. During training, the loss objective (L) is such that the generator maximizes the discriminator's classification loss, whereas the discriminator's objective is to minimize that loss [17]:

$$\min_G \max_D L[D,G], where\ L[D,G] = E_{x\sim data}[log(D(x))] + E_{z\sim p_z(z)}[log(1-D(G(z))]$$

In the styleGAN architecture, the latent space vector $\mathbf{z}$ is mapped via a multilayer-perceptron mapping network to an intermediate vector

space **w**. This vector **w** is transformed into so-called styles via learned affine transformations. Styles are then added to intermediate layers of the generator via adaptive instance normalization (AdaIn) [33] (Fig. 1A). styleGAN2 reuses the discriminator architecture of Progressive GANs by Karras et al. [39]. Detailed training parameters can be found in Appendix A.

2.2.2. Latent diffusion models

We implement and test three different variants of latent diffusion models [25]. Two models are trained from scratch on histology data with two different autoencoders: a Kullback-Leibler KL-autoencoder [40], referred to as KLF8-DM, and a vector quantization VQ-autoencoder [41], referred to as VQF8-DM. Thirdly, we finetune the released stable diffusion model v1.4 trained on the LAION-400 M dataset [42].

In DMs, Gaussian noise is iteratively added to the original image according to a variance schedule with a large total number of steps (T = 1000), such that the distribution of pixel values in the resulting images tends towards a normal distribution. The diffusion network ϵ_θ is trained by minimizing the loss objective L_{DM} that resembles denoising score matching [43] for each time step t, $t \epsilon N, t \leq T$ [18]:

$$L_{DM} = E_{x,\epsilon\sim N(0,1),t}\left[||\epsilon - \epsilon_\theta(x_t t)||_2^2\right]$$

The learned function ϵ_θ can subsequently be used for a stepwise denoising process: Starting from an image with each pixel initialized by a normal random distribution, an image can be generated by successive application of denoising kernels that are modeled by ϵ_θ. Image generation in DMs is very resource intensive, and to speed up image generation, we use the denoising diffusion implicit models DDIM-sampler with 200 denoising steps throughout [44].

In LDMs, the diffusion process is only performed in a latent space with a reduced number of dimensions (Fig. 1B). The autoencoder encodes original images into this latent space. Vice versa, generated images need to be decoded from the latent space. In this analysis, spatial dimensions are reduced by a factor of eight from 256 × 256 pixels to 32 × 32 pixels. This choice results in a good trade-off between computational efficiency and obtaining perceptually faithful results [25].

For the LDMs in this paper, the denoising network is a time-conditional U-Net [45] that is augmented by a cross-attention mechanism to allow for class-conditioned image synthesis [25]. In stable diffusion, class conditioning is achieved via text strings tokenized by the BERT-tokenizer [46] and subsequently embedded via a transformer [47]. Embedded inputs are then used in the cross-attention mechanism (Fig. 1B). For the KLF8-DM and VQF8-DM models, class embeddings of 512 output dimensions are obtained from binarized class labels via a linear layer [25].

Training a LDM from scratch is a two-stage process: First, training the autoencoder, and second, training the DM in latent space. Using the autoencoder, intermediate results of the diffusion process can also be decoded from latent space to 256 × 256 images to examine the image generation process. Detailed training parameters can be found in Appendix B.

2.3. Code availability

Our implementations are based on publicly available repositories at GitHub: https://github.com/CompVis/latent-diffusion, GitHub: https://github.com/CompVis/stable-diffusion, and GitHub: https://github.

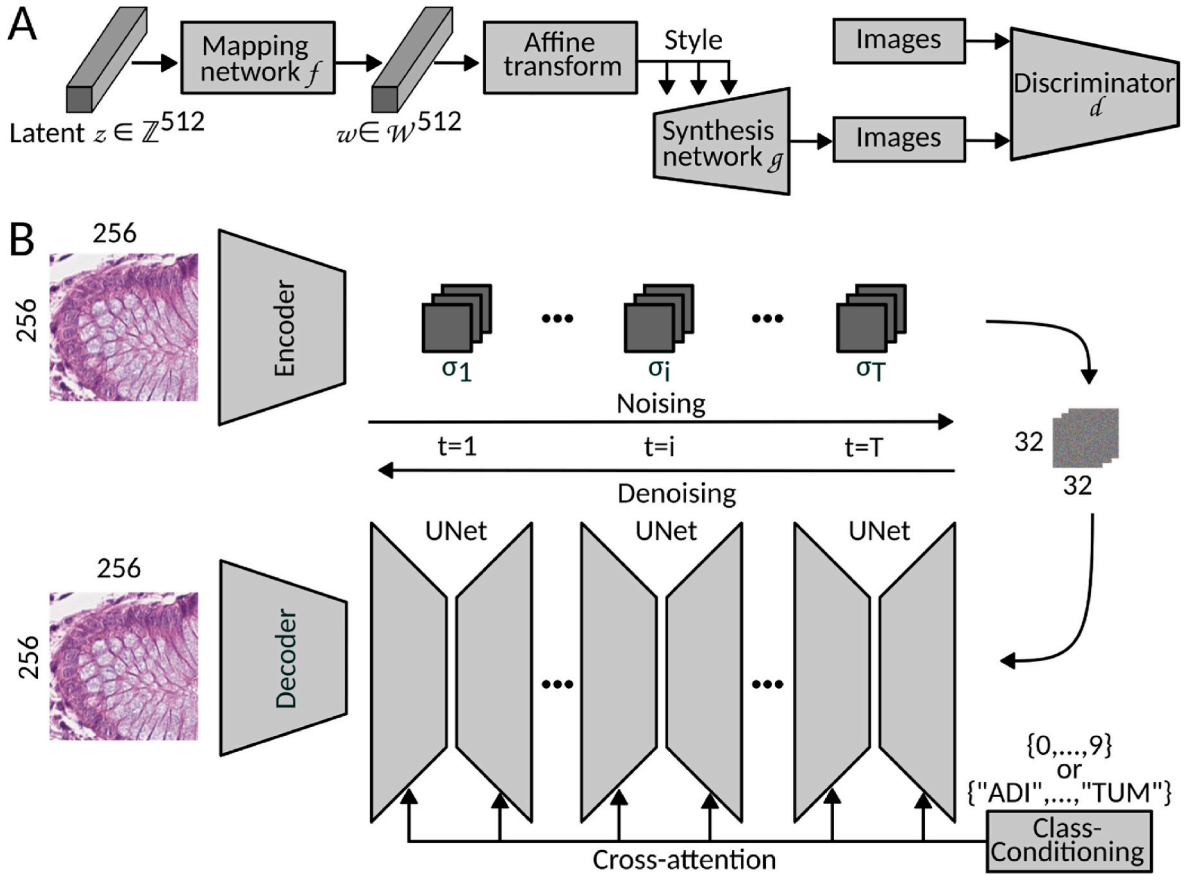

Fig. 1. Overview of model architectures. A: Overview of the styleGAN2 model architecture. A latent vector z is transformed into a style vector that is passed to the synthesis network g. The discriminator network d subsequently classifies generated and real images. **B**: Overview of the LDM architecture. The encoder compresses an image to latent space that is subsequently perturbed by Gaussian-noise kernels σ for T steps. In the denoising process, the added noise is predicted and the input image is reconstructed.

com/NVlabs/stylegan3. Our trained models, configuration files with parameter setups, and code extensions to the official latent diffusion repository are available at GitHub: https://github.com/janniehues/my_LDM and Zenodo: 10.5281/zenodo.10838706.

2.4. Evaluation metrics

In LDMs, the quality of the deployed autoencoders is assessed by the Multiscale Structural Similarity Index Measure (MS-SSIM) [48] between input and output image pairs (Fig. 2A). Output images are obtained by encoding the input image and decoding the output image using the autoencoder. The average MS-SSIM is calculated across the full dataset for each autoencoder.

To evaluate the quality of generated images via all generative models, we use three different metrics:

First, we utilize the Frechet inception distance (FID) [35] to measure the similarity between images in the training dataset and synthesized images. To compute the FID we forward all images through an ImageNet pre-trained Inception-v3 model [49], and use the resulting feature distributions in each set to calculate respective means (μ, $\widehat{\mu}$) and covariance matrices ($\Sigma, \widehat{\Sigma}$) to compute:

$$FID = |\mu - \widehat{\mu}|_2^2 + Tr\left(\Sigma + \widehat{\Sigma} - 2(\Sigma\widehat{\Sigma})^{1/2}\right)$$

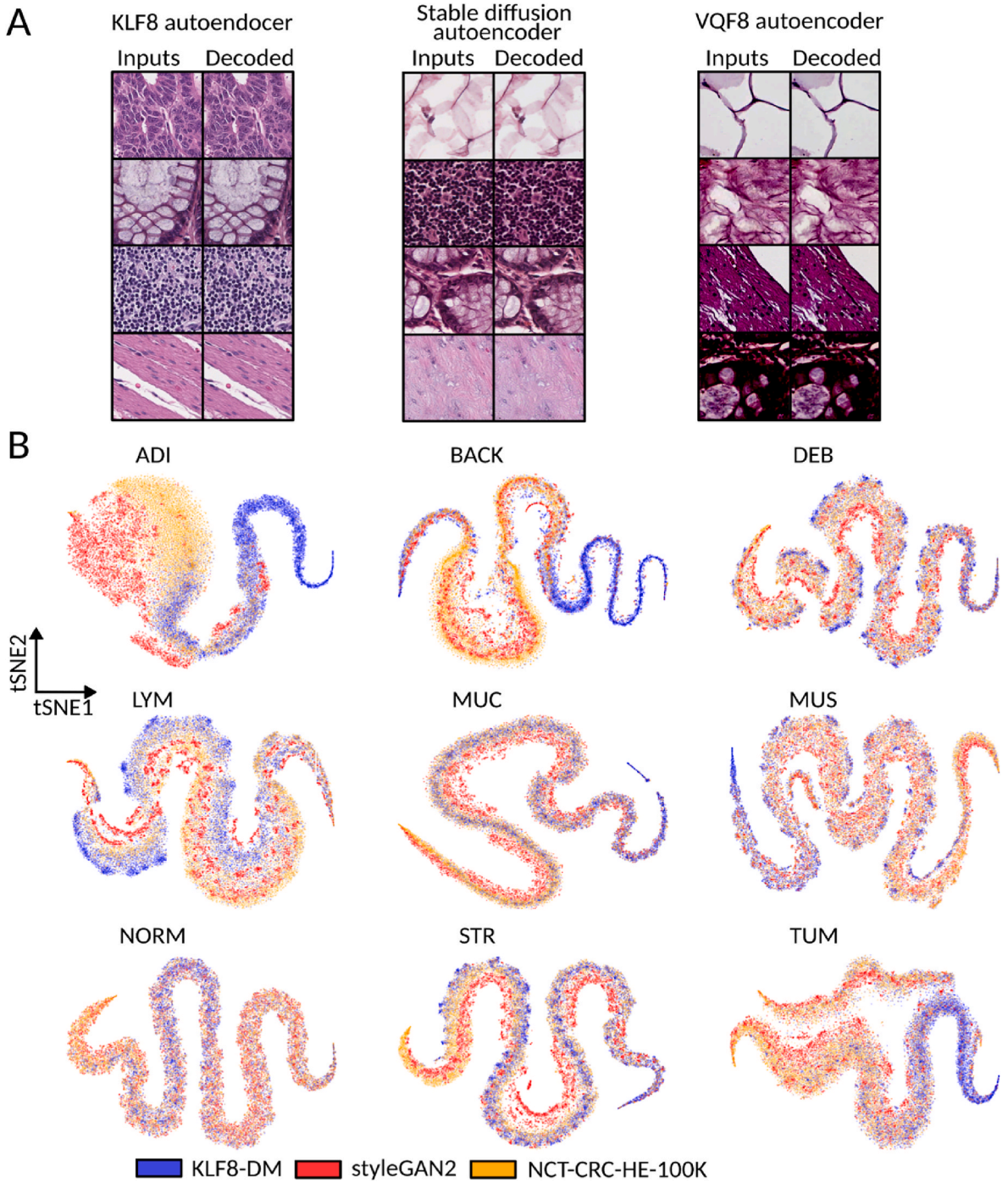

Fig. 2. Autoencoder results, and t-SNE plots. A: Input examples and decoded images for the KLF8 autoencoder (left), stable diffusion's autoencoder (middle), and the VQF8 autoencoder (right). All three autoencoders provide a very good reconstruction quality.

A lower FID score indicates that the images in the training and generated sets are more similar, and a higher FID score indicates that images are less similar to each other. In our analysis, FIDs are calculated using the function as implemented in the clean-FID package [50]. To evaluate a generative method's performance, the FID score for any class is calculated between 10,000 generated images and all images in the training dataset for a particular class.

Second, we plot t-SNE distributions from all generated images by the KLF8-DM and styleGAN2 models together with images from the NCT-CRC-HE-100K dataset for each data class. To this end, all images in a dataset are passed through the ImageNet-pre-trained Inception-v3 model to obtain feature vectors. These feature vectors are used to fit t-SNE distributions and resulting plots are qualitatively interpreted (Fig. 2B).

Finally, three medical experts rate the synthesized images' quality to obtain expert rating scores: This rating is performed separately for every tissue class. For each class, ten images are selected randomly from all four trained models and combined with ten randomly selected images from the training dataset, totaling 50 images per class. The experts are then asked to distinguish real from synthesized images in this combined set (Fig. 3). The difference in the fraction of synthetic images and of real images classified as real images averaged across the experts defines the expert rating score for each class and generative model:

$$\textit{Expert rating score}\ (k,m) = \frac{1}{\mathrm{N_{Experts}}} \sum_i \left(\frac{\mathrm{n}^{(i)}_{\mathrm{g,k,m}}}{\mathrm{N_{tot}}} - \frac{\mathrm{n}^{(i)}_{\mathrm{r,k}}}{\mathrm{N_{tot}}} \right)$$

where $n^{(i)}_{g,k,m}$ and $n^{(i)}_{r,k}$ is the number of generated and real images, respectively, classified as real for expert i, tissue class k and model m. $N_{Experts} = 3$, is the number of experts, and $N_{tot} = 10$, is the total number of images for each model and class. Apriori, the medical experts did not know anything about the composition of the image sets given to them. In total, each expert evaluated 450 images.

B: t-SNE distributions for generated images for the KLF8-DM (blue) and styleGAN2 (red) model together with the distribution for images from the NCT-CRC-HE-100K training dataset (orange) for each tissue class. The t-SNE distributions show that styleGAN2 images tend to form clusters, whereas this is not the case for the KLF8-DM. The t-SNE distributions for the KLF8-DM model are better aligned with the distributions of the training data.

2.5. Evaluation of data memorization

To evaluate if an image in a synthetic data set is memorized from the training dataset we follow and modify the procedure by Akhbar et al. [30]. In this analysis, we randomly select 200 images from the generated dataset and rotate each by 90, 180, and 270°. For each selected image, we calculate the pixel-wise correlation coefficient with all images in the training dataset and select the highest correlation coefficient. The pixel-wise correlations are evaluated on grayscale images to reduce computational costs and to only highlight structural similarities between the images. Finally, the distribution of the correlation coefficient of the highest correlated image pairs is plotted for every class (Fig. 4, Table 1). In this way, data memorization in synthetic datasets is visualized and can be compared between different generative methods.

2.6. Classifier training

To assess the impact of augmenting the training dataset with

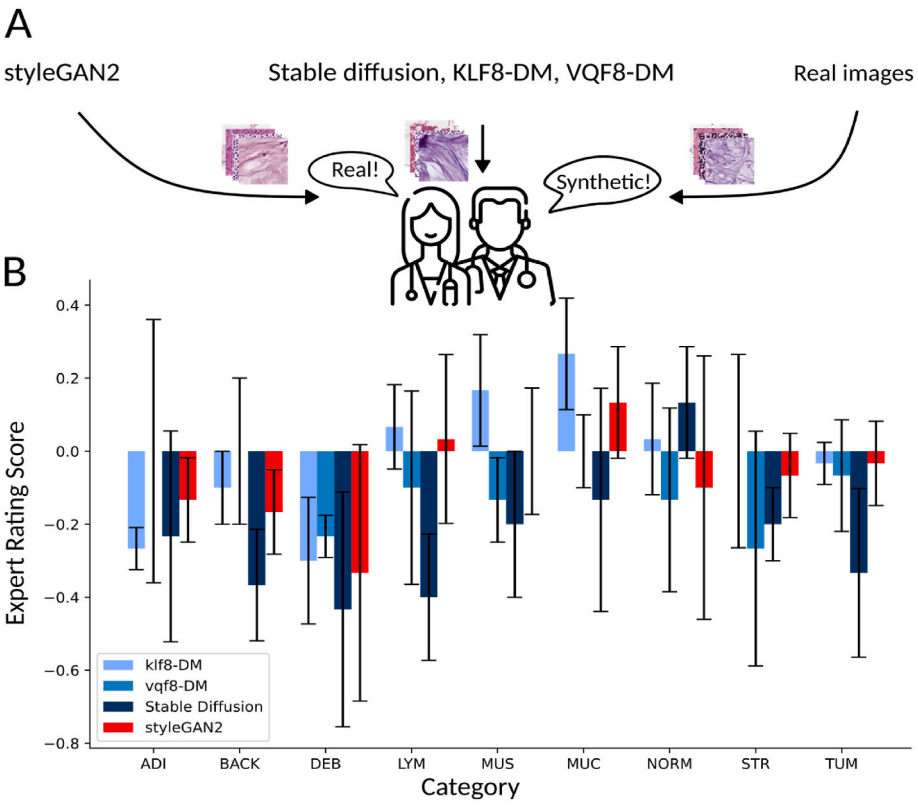

Fig. 3. User study sketch and results of user study. A: Sketch of the user study. Generated images from all four models and real images are presented to three medical experts without revealing the images' origins. Experts rate images either as generated (0) or real (1). B: Expert rating score results. A score of zero means that generated images are as often rated "real" as genuinely real images are rated "real". A positive score can be interpreted as the fraction of synthetic images that are more often rated "real" than real images are rated "real", and vice versa for negative scores. Error bars give the standard deviation across the three expert ratings. We find that pathologists can still distinguish synthetic from real images. Latent diffusion models using the KLF8 encoder (KLF8-DM), however, achieve the highest expert rating scores for most categories, which means that they can be least distinguished by pathologists.

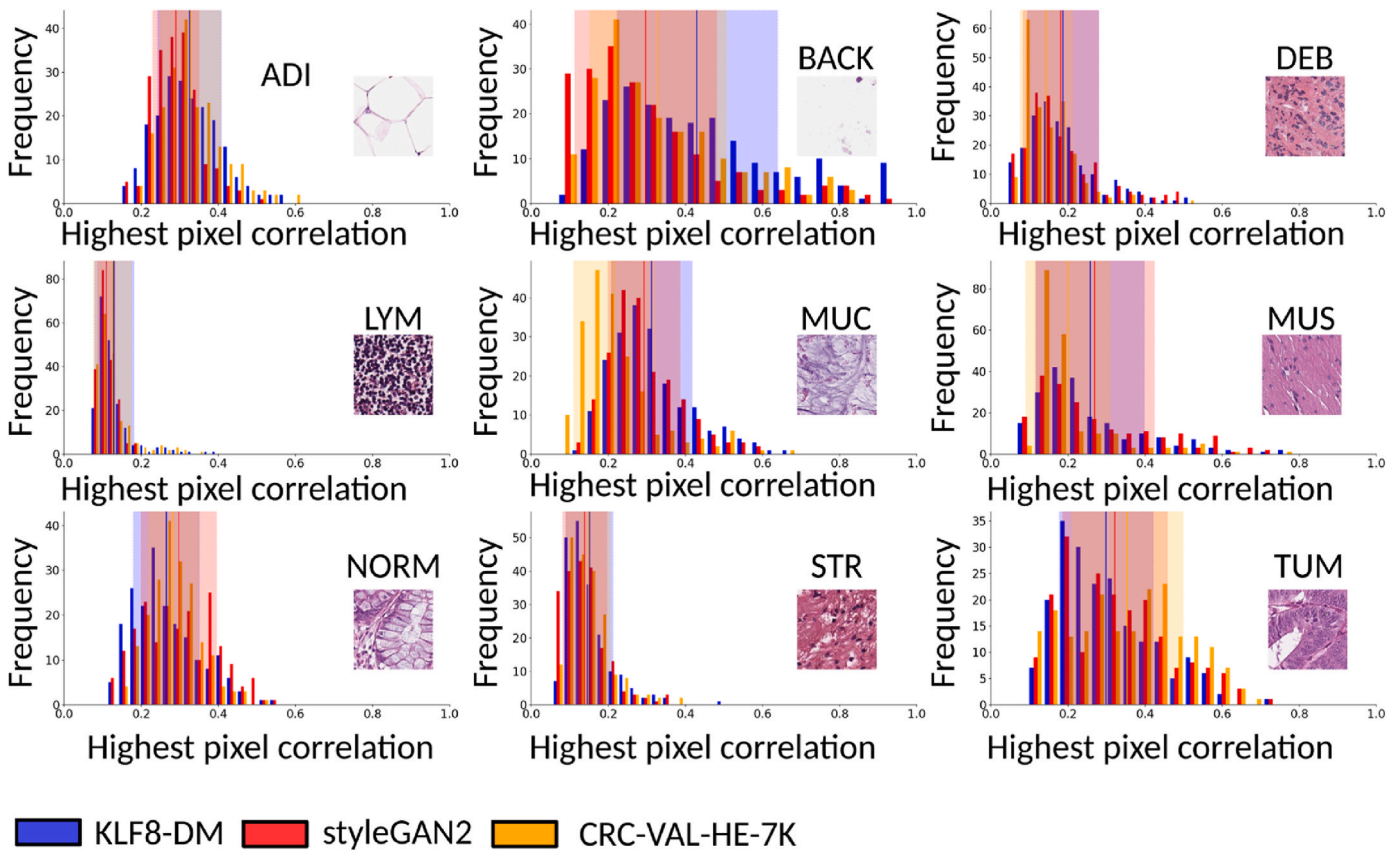

Fig. 4. Highest correlation coefficients for pixel-wise correlations between generated images and training images: For each tissue class, the distribution of the correlation coefficients is shown for the highest correlated pair of images from the NCT-CRC-HE-100K training dataset with images from the KLF8-DM (blue), the styleGAN2 (red), and the independent CRC-VAL-HE-7K test set (orange). A pixel correlation value of one for any generated image means that the model has seen an identical image during training. A pixel correlation of zero means that no similar configured image was seen during training. The test dataset is independent of the training dataset and serves as an indicator of expected random correlations for each class. The vertical bands and shaded areas show the mean ± standard deviation for individual pixel correlation distributions. Across the classes, the distributions of the highest correlation coefficients are very similar for the three datasets, and the mean values are within the range of one standard deviation from each other. This shows that negligible data memorization occurs in the KLF8-DM model.

Table 1
Correlation statistics to assess the degree of data memorization for generative models. Correlations are calculated between images generated by the KLF8-DM and styleGAN2 models, and for images from the CRC-HE-VAL-7K dataset with images from the training dataset. For each tissue class the 5%- and 95%-percentiles, mean and median values are shown for correlations of highest correlation image pairs. The degree of memorization is similar for all assessed datasets in each tissue class.

ADI	mean	median	p5	p95	MUS	mean	median	p5	p95
KLF8-DM	0.33	0.32	0.2	0.46	**KLF8-DM**	0.26	0.21	0.10	0.55
styleGAN2	0.29	0.29	0.21	0.4	**styleGAN2**	0.27	0.22	0.09	0.58
CRC-VAL-HE-7K	0.33	0.31	0.22	0.47	**CRC-VAL-HE-7K**	0.20	0.16	0.12	0.45
BACK	**mean**	**median**	**p5**	**p95**	**NORM**	**mean**	**median**	**p5**	**p95**
KLF8-DM	0.43	0.38	0.17	0.84	**KLF8-DM**	0.27	0.25	0.16	0.42
styleGAN2	0.30	0.25	0.09	0.75	**styleGAN2**	0.30	0.29	0.15	0.47
CRC-VAL-HE-7K	0.33	0.28	0.12	0.70	**CRC-VAL-HE-7K**	0.28	0.28	0.19	0.39
DEB	**mean**	**median**	**p5**	**p95**	**STR**	**mean**	**median**	**p5**	**p95**
KLF8-DM	0.19	0.17	0.07	0.38	**KLF8-DM**	0.15	0.14	0.09	0.26
styleGAN2	0.18	0.15	0.06	0.39	**styleGAN2**	0.14	0.13	0.07	0.24
CRC-VAL-HE-7K	0.14	0.13	0.07	0.27	**CRC-VAL-HE-7K**	0.15	0.14	0.08	0.25
LYM	**mean**	**median**	**p5**	**p95**	**TUM**	**mean**	**median**	**p5**	**p95**
KLF8-DM	0.13	0.12	0.09	0.24	**KLF8-DM**	0.30	0.27	0.14	0.53
styleGAN2	0.11	0.10	0.08	0.15	**styleGAN2**	0.32	0.30	0.14	0.57
CRC-VAL-HE-7K	0.13	0.11	0.08	0.25	**CRC-VAL-HE-7K**	0.35	0.36	0.13	0.58
MUC	**mean**	**median**	**p5**	**p95**					
KLF8-DM	0.31	0.29	0.18	0.53					
styleGAN2	0.29	0.27	0.17	0.47					
CRC-VAL-HE-7K	0.22	0.19	0.11	0.45					

synthetic images on multiclass classifier performance, we train a ResNet34 [51] on three training datasets: the NCT-CRC-HE-100K dataset, the synthetic KLF8-DM dataset alone, and the combination of the two datasets. Before training, the non-normalized tiles are normalized using the Macenko algorithm [11]. The training pipeline is implemented using the fastai software package [52]. The network is trained to classify tissue tiles into correct tissue classes. Detailed training parameters can be found in the Appendix. The classifiers' performances are evaluated via classification accuracy across all tissue classes in the CRC-VAL-HE-7K test dataset.

2.7. Ethics statement

This study was carried out in accordance with the Declaration of Helsinki. The analysis was approved by the ethics committee of the Medical Faculty of Technical University of Dresden (BO-EK-444102022). No individual patient consent was required for this retrospective analysis of anonymized data.

3. Results

3.1. All autoencoders achieve a high image reconstruction quality

We found that generated images are of high quality independent of the generative model used (Fig. 5A). When generating images using LDMs, coarse structures are initially generated that are followed by the addition of finer details in later stages (Fig. 5B). In Fig. 5, the images are obtained by decoding the intermediate results of a 1000-step diffusion process from latent space to the original image space. First, we evaluate the reconstruction quality of the autoencoder: A perfect autoencoder can encode and subsequently decode an input image without any reconstruction loss (Fig. 5A). For the three deployed autoencoders, the KLF8-autoencoder obtains an MS-SSIM score of 91.6 ± 5.3%, stable diffusion's pre-trained KLF8-autoencoder a score of 88.3 ± 7.0%, and the VQF8-autoencoder a score of 86.4 ± 7.8% across all tiles in the training dataset, all within the range of one standard deviation.

Together, the samples and the evaluation demonstrate the high level of reconstruction quality of the employed autoencoders and that all architectures can be used to generate high-quality images in LDMs with a slight preference towards the KLF8-autoencoder as shown by the higher MS-SSIM.

3.2. The KLF8-DM model outperforms all other models across the evaluation metrics

For each class and model, we assess the similarity of generated images and images in the training dataset by FID scores (Table 2). For the majority of complex tissue classes, the KLF8-DM scores the lowest (best) FIDs throughout outperforming all other generative models (FID(DEB) = 10.6, FID(MUC) = 15.3, FID(MUS) = 17.1, FID(NORM) = 14.3, FID(STR) = 12.4, FID(TUM) = 10.3). For the lymphocytes class, all models perform equally well (KLF8-DM: FID(LYM) = 22.8). For the less complex

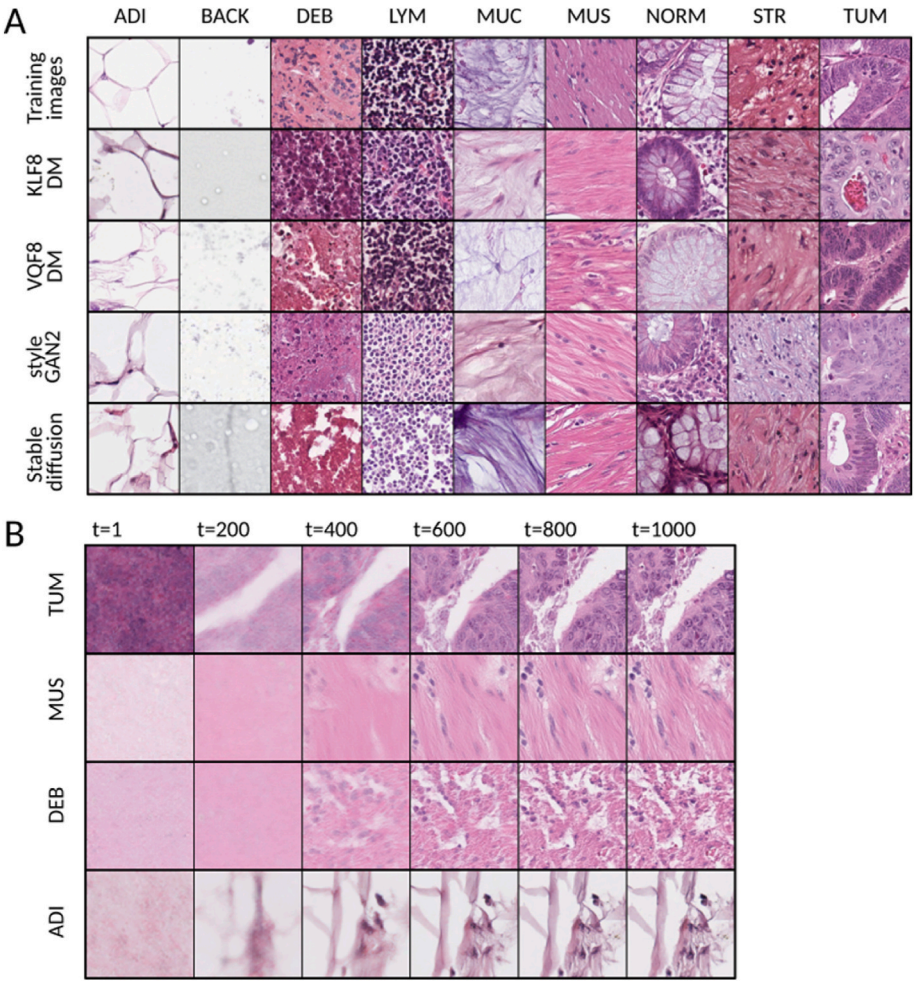

Fig. 5. Example images and progressive rows for LDM image generation showing image status after t time steps. A: Examples of generated and real images for each category and method. B: Rows show generated images after different diffusion steps. Images are shown after 1, 200, 400, 800, and 1000 diffusion steps. Results are shown for tumor, muscle, debris, and adipose classes.

Table 2
FID scores by method and tissue category. Latent diffusion using the KLF8 encoder systematically outperforms all other approaches in terms of FID score for the generation of more complex tissue types mucus (MUC), muscle (MUS), normal stroma (NORM), stroma (STR), debris, and tumor tissue (TUM). styleGAN2 obtains the best FID score for lymphatic tissue (LYM). Latent diffusion using the VQF8 encoder obtains the best results for the generally less complex fat and background tissues.

	ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM
Latent diffusion (KLF8)	73.2	73.1	**10.6**	22.8	**15.3**	**17.1**	**14.3**	**12.4**	**10.3**
Latent diffusion (VQF8)	37.2	**13.2**	18.6	18.8	27.0	23.0	24.8	21.9	21.4
Fine-tuned stable diffusion	105.9	164.6	24.5	21.6	16.7	23.6	17.8	20.0	21.4
styleGAN2	**32.9**	36.8	15.0	**18.7**	22.4	20.6	18.9	15.6	15.5

adipose and background classes, the VQF8-DM and the styleGAN2 models outperform the other two LDMs. The styleGAN2 model and the VQF8-DM score lowest for the adipose (FID(ADI) = 32.9) and background (FID(BACK) = 13.2) classes, respectively.

Further, we compare the t-SNE distributions between generated images from an LDM and the styleGAN2 model. Choosing the KLF8-DM as a representative of the LDMs as it has the lowest FIDs for the more complex classes, the t-SNE distributions for the different classes show that images from the KLF8-DM model are more aligned with the training data than styleGAN2 images (Fig. 2B). This is in particular true for the DEB, LYM, MUC, STR, and TUM tissue classes. For these tissue classes the styleGAN2 images form clusters outside the distribution obtained by corresponding images in the NCT-CRC-HE-100K training dataset. This happens only to a lesser extent for images generated by the KLF8-DM model. We note that the embeddings produced by t-SNE are anisotropic, permitting only qualitative analyses.

Finally, expert ratings are used to evaluate generated images for each class and model (Fig. 3). The KLF8-DM obtains the highest expert rating scores for lymphocyte, cancer-associated stroma, muscle, mucus, and tumor classes, but obtains the lowest scores for the adipose class. The VQF8-DM scores the highest for background, adipose, and cell debris classes. styleGAN2 obtains the highest expert rating score for the tumor class. Finally, fine-tuned stable diffusion scores the highest for normal mucosa, and the lowest for background, debris, lymphocyte, muscle, mucus, and tumor classes. The inter-rater agreement as measured by Fleiss'-Kappa [53] was slightly positive across images from all generative models, and medical experts agreed stronger on the classification of real images than of generated images (average Fleiss's-Kappas: KLF8-DM: 0.04, VQF8-DM: 0.08, Stable Diffusion: 0.02, styleGAN2: 0.06, real images: 0.22).

Together, these data show that the KLF8-DM generates images of the highest FID as well as expert rating scores for more complex tissue classes, but is inferior in generating images for less complex tissue classes. The fine-tuned stable diffusion model also shows a better performance for more complex than less complex tissue classes, however, generated images have higher FID scores for eight of the nine categories. Also, images from fine-tuned stable diffusion are the least likely to be identified as authentic across the nine classes. The styleGAN2 model and VQF8-DM are both better at generating images for less complex adipose and background tissue classes. This is true in terms of FID as well as expert rating scores. The expert rating scores, however, come with large standard deviations such that expert ratings are compatible with the scores of real images for most classes and models. Thus, medical experts have difficulties distinguishing between real and synthetic images. This shows that all models generate authentic images of high quality, without displaying striking artifacts. Finally, the t-SNE distributions furthermore show that the generated images from the KLF8-DM are more similarly distributed to the images in the training dataset than images from the styleGAN2 model. Assuming that the more complex the images in a particular class are, the more relevant this class is for training an image classifier, we identify the KLF8-DM as the best-performing model in terms of FID and expert rating score for data augmentation.

3.3. Data memorization is not observed in KLF8-DM and styleGAN2 models

We evaluate data memorization in the generated image datasets by calculating pixel-wise correlations of a subset (n = 200) of grayscale images with all grayscale images from the NCT-CRC-HE-100K dataset. This analysis is performed on generated images from the styleGAN2 and the KLF8-DM models. The analysis is also repeated for the images in the CRC-VAL-HE-7K test dataset and this serves as an indicator of random correlations among images in a particular tissue class. The distribution of the highest correlation coefficient is very similar for both images from the styleGAN2 and the KLF8-DM models across all nine tissue classes (Fig. 4, Table 1). In particular, the averages for the highest correlation coefficients for a particular distribution lie within the range of a standard deviation from the average of the corresponding distribution of the other model for a particular class. This is also true when comparing the distributions for any of the two generating models to the distribution of the highest correlation coefficients obtained for images from the CRC-VAL-HE-7K test dataset. Together these data show that data memorization is not observed in either set of generated images for the styleGAN2 or the KLF8-DM models.

3.4. Dataset augmentation using synthetic images improves classifier performance

We train a classifier on three sets of images: images generated by the KLF8-DM, Macenko-normalized images from the NCT-CRC-HE-100K dataset, and the combination of the two datasets. The performances of trained classifiers are evaluated on the Macenko-normalized tiles in the separate CRC-VAL-HE-7K test dataset (Table 3). For classifiers trained on individual datasets, adipose, background, normal mucosa, and tumor classes are classified with at least 95% accuracy. Debris, lymphocytes, mucus, muscle, and stroma classes are more difficult to classify, with classification accuracies ranging from 43% to 90%. The overall accuracy across all classes is similar between classifiers trained on the KLF8-DM generated images and NCT-CRC-HE-100K images with 80% and 81%, respectively. Combining the two datasets into one single training dataset results in a classifier that considerably improves the overall classification performance. For any more difficult class, the resulting classifier is either much more accurate than either of the classifiers trained on the individual datasets (STR, LYM) or close to the accuracy of the classifier trained on the NCT-CRC-HE-100K alone (DEB, MUC). By augmenting the training dataset with synthetic images in this way, the overall classification accuracy on the test set can be improved by 4% from 81% to 85%. This improvement in classifier performance when adding synthetic images to the dataset is in agreement with similar findings obtained by Ye et al. [32].

Together these results indicate that LDMs can be used as an augmentation technique to improve the performance of histopathology image classifiers. Also, the performance of the classifier trained on images from the LDM alone is comparable to the performance of the classifier trained on real images across all classes. This shows that the image quality achieved by LDMs is sufficient for training DL architectures such that LDMs can be used for data sharing and to address the problem of class imbalances.

Table 3
Accuracy scores for tissue classifiers. Tissue classifiers are trained on images generated by the KLF8-DM, the training dataset images, and the combined dataset of the two. Classification performances using a combination of both datasets are superior to classification from a single dataset alone.

	ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM	AVERAGE
KLF8-DM	**0.99**	**1.**	0.43	0.74	**0.78**	0.78	**0.99**	0.57	**0.95**	0.80
real images	0.95	0.97	**0.90**	0.71	0.63	**0.80**	**0.99**	0.44	**0.96**	0.81
combined	0.95	**1.**	**0.89**	**0.88**	0.62	0.77	**0.99**	**0.62**	**0.96**	**0.85**

4. Discussion

Generating synthetic histopathological images is relevant in multiple aspects: Generative models are used for educational purposes [7], and can help resolve issues when applying DL to automate decision-making processes in histopathological workflows. These issues include patient privacy concerns, class imbalances in many histopathological datasets, as well as mitigated classifier performances for biomarker prediction on previously unseen data. Generative models may help to address these problems: First, generated images can augment existing datasets, making trained models less prone to overfitting on training data. Second, class imbalances are corrected by presenting DL algorithms with characteristic generative data for any scarce class. In the case of GANs, Krause et al. [14] showed that training on synthetic images results in superior image classifiers of microsatellite instability in CRC compared to classifiers trained on real images alone. Their work proved the enhanced robustness of resulting classifiers when using generated images to augment training datasets.

In this paper, we examine a similar potential of LDMs for augmenting image datasets in histopathology and compare the results to a state-of-the-art GAN for various metrics in a systematic way. Previous works have already shown the superiority of DM compared to GANs outside and inside the medical domain [20,21,23]. For medical data, however, data privacy is of the utmost priority. Akbar et al. [30] and Dar et al. [31] showed that data memorization can occur in DM. Therefore, we combine the quantitative comparison of synthetic images with an assessment of data memorization.

In our analysis, all evaluated generative DL methods produce images of high quality across the tissue classes, and medical experts have difficulties distinguishing original from generated images. A result that has already been observed by Krause et al. with GANs [14]. Images generated by the KLF8-DM are of the highest quality (expert rating scores) and more similar to real images (FID score) in more complex tissue classes than images from the other models. For the less complex adipose and background classes, the fine-tuned stable diffusion and the KLF8-DM, are outperformed by the VQF8-DM in terms of FID and expert rating scores. This result shows that the former two, both using the KLF8-autoencoder, have problems in generating predominantly white images. Depending on the complexity of histopathological classes, it might therefore be favorable to use the VQF8-autoencoder instead. Furthermore, fine-tuning the stable diffusion model using the same training setup as the KLF8-DM results in lower-quality images. As the diffusion process in latent space is the same, this must be attributed to the difference in deployed autoencoders. This can be due to two reasons: First, the deployed autoencoder's reconstruction quality is different, and second, histopathological images are encoded to less favorable feature maps for the diffusion process to take place. Both autoencoders show a high reconstruction quality with MS-SSIM scores of 88.3 $\pm$ 7.0% and 91.6 $\pm$ 5.3%, respectively (Fig. 2A). Therefore, the lower image quality in fine-tuned stable diffusion images could be due to the interplay of the autoencoder and the diffusion process in latent space.

We identify the KLF8-DM as our strongest generative LDM across the classes. Taking the KLF8-DM as a representative for LDMs, we further find that images from the KLF8-DM and the styleGAN2 model form different patterns on t-SNE distributions. T-SNE distributions of KLF8-DM images spread out similarly to real images. In contrast, styleGAN2 images tend to form clusters on the t-SNE plots which is not observed for real images (Fig. 2B). Our findings across the various metrics are in line with previous findings of LDMs beating GANs within the medical domain [21].

Using the KLF8-DM, we find that the overall accuracy of classifiers trained on generated images is similar to the accuracy of a classifier trained on real images alone. This result is promising and confirms the findings by Krause et al. [14]. It is a proof of concept that data can be shared via LDMs to train DL architectures either to account for dataset imbalances or to circumvent patient privacy issues. For the latter aspect, however, data should not be memorized by the models. This is not the case for the KLF8-DM, which we test by evaluating pixel-wise correlations between generated images and images in the training dataset. Data memorization should always be assessed before sharing any LDM that was trained on sensitive data. In the medical domain, it would be desirable to establish a standard for data memorization checks in the future. Finally, we obtain a superior classifier by augmenting the real dataset with the synthetic data. This is an important result as it shows that LDMs can be safely and successfully applied in the medical domain for downstream classification tasks.

All models are trainable in a straightforward way. During the training of the styleGAN2 model, however, mode collapse was encountered several times. When mode collapse occurs, the model's output collapses to a few modes or patterns to produce repetitive and unrealistic outputs [54,55]. In this case, the training has to be restarted from the last non-collapsed training point. We did not observe this phenomenon for diffusion models. Training the styleGAN2 took a couple of days for one training run, as training had to be restarted multiple times due to mode collapse. Ignoring mode collapse, the total consecutive training time for the styleGAN2 and the LDMs was similar, taking 8h and 24h on two NVIDIA RTX A6000, respectively. The advantage of GAN models is that image synthesis is fast. Ten thousand images can be generated within minutes. For latent diffusion processes, however, this process takes 10 to 20 times longer. During inference for a diffusion model on an NVIDIA RTX A6000, the generation of a single image takes 0.7s, totaling approximately 2h for the generation of ten thousand images for one class. Generated images for all models can be found under Zenodo: 10.5281/zenodo.10838706.

A few limitations of this study exist: First, the analysis is performed on a dataset only containing tissue classes, and not on a dataset including biomarker-typical morphologies. Second, the training data is far from being representative of the full histopathological domain. With the advent of foundational models in histopathology, we may have foundational generative models soon to enrich datasets. Such a foundation model requires more diverse data than used in our analysis. Our results indicate such a model may be used to improve DL algorithm performances. Third, this analysis was carefully conducted, however, it can not be excluded that other training hyperparameters exist that result in better models. Fourth, we only checked a limited number of autoencoder architectures and focused on the most promising autoencoders tested in the non-medical domain [25]. Fifth, we only generated images of histopathological tiles. In contrast, modern DL biomarker classification algorithms usually analyze histopathological whole slide images (WSIs). The generation of such WSIs is in principle possible, but lies beyond the scope of this paper. Finally, we used an ImageNet-pre-trained Inception-v3 model to extract features for downstream tasks. This might not be an optimal choice for an evaluation of histopathological images.

5. Conclusion

This paper serves as an evaluation of LDMs in histopathology. We trained different LDMs and compared the resulting image qualities to images from a styleGAN2 model. The KLF8-DM model gives the best LDM and outperforms styleGAN2 in different evaluation metrics (FID, t-SNE, expert rating scores). We show that images generated by the KLF8-DM are not memorized from the training data. Augmenting the NCT-CRC-HE-100K datasets by KLF8-DM's generated data for training, results in a superior tissue classifier as compared to training on the real dataset alone. This shows that LDMs can be used to share data and augment existing datasets in a patient-privacy-respecting way to improve downstream classification tasks.

In future work, images from LDMs may be used to augment datasets for state-of-the-art biomarker prediction, and the impact on resulting classifier performances should be evaluated. State-of-the-art biomarker prediction, however, usually relies on the classification of histopathological WSIs. Hence the generation of WSIs via LDMs is a missing step and remains to be addressed. The ultimate goal is to train a generative foundation model for histopathology. This model could then generate images across the full span of histopathological modalities. By applying LDMs to improve the accuracy of state-of-the-art DL algorithms in histopathology, LDMs could directly impact clinical DL applications.

Funding

JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111), the Max-Eder-Programme of the German Cancer Aid (grant #70113864), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (Transplant.KI, 01VSF21048) the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312) and the National Institute for Health and Care Research (NIHR, NIHR213331) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. S.J.W. was supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS" and the Add-on Fellowship of the Joachim Herz Foundation. DT is supported by the German Federal Ministry of Education (TRANSFORM LIVER, 031L0312A; SWAG, 01KD2215B) and the European Union's Horizon Europe and innovation programme (ODELIA, 101057091). ATP reports effort support via grants from NIH/NCI U01-CA243075, NIH/NIDCR R56-DE030958, NIH/NCI R01-CA276652, SU2C (Stand Up to Cancer) Fanconi Anemia Research Fund – Farrah Fawcett Foundation Head and Neck Cancer Research Team Grant, European Union Horizon Program (I3LUNG).

CRediT authorship contribution statement

Jan M. Niehues: Conceptualization, Software, Writing – original draft, Formal analysis. **Gustav Müller-Franzes:** Conceptualization, Writing – review & editing. **Yoni Schirris:** Conceptualization, Writing – review & editing. **Sophia Janine Wagner:** Conceptualization, Visualization, Writing – review & editing. **Michael Jendrusch:** Conceptualization, Writing – review & editing. **Matthias Kloor:** Conceptualization, Writing – review & editing. **Alexander T. Pearson:** Conceptualization, Writing – review & editing. **Hannah Sophie Muti:** Conceptualization, Investigation, Writing – review & editing. **Katherine J. Hewitt:** Conceptualization, Investigation, Writing – review & editing. **Gregory P. Veldhuizen:** Conceptualization, Investigation, Writing – review & editing. **Laura Zigutyte:** Conceptualization, Writing – review & editing. **Daniel Truhn:** Conceptualization, Supervision, Writing – review & editing. **Jakob Nikolas Kather:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

JN Kather reports a relationship with AstraZeneca Pharmaceuticals LP that includes: consulting or advisory and speaking and lecture fees. JN Kather reports a relationship with Bayer AG that includes: speaking and lecture fees. JN Kather reports a relationship with Eisai Inc that includes: speaking and lecture fees. JN Kather reports a relationship with Merck Sharp & Dohme UK Ltd that includes: speaking and lecture fees. JN Kather reports a relationship with Bristol-Myers Squibb Company that includes: speaking and lecture fees. JN Kather reports a relationship with F Hoffmann-La Roche Ltd that includes: speaking and lecture fees. JN Kather reports a relationship with Pfizer Inc that includes: speaking and lecture fees. JN Kather reports a relationship with Fresenius Kabi Germany that includes: speaking and lecture fees. JN Kather reports a relationship with Owkin France that includes: board membership and speaking and lecture fees. JN Kather reports a relationship with Panakeia, UK that includes: board membership and speaking and lecture fees. JN Kather reports a relationship with DoMoreDiagnostics that includes: board membership and speaking and lecture fees. JN Kather reports a relationship with Histofy, UK that includes: speaking and lecture fees. JN Kather reports a relationship with StratifAI GmbH that includes: board membership and equity or stocks.

Acknowledgments

None.

Appendix

A. Training Details: styleGAN2

We train our styleGAN2 model with **z** and **w** dimensions of 512. The loss function is regularized by R1 loss regularization [56] with $\gamma = 0.5$, a style-mixing probability of 0.9, and path length regularization of weight 2 [33,57]. The network weights are updated using the Adam optimizer [58] ($\beta_1 = 0$, $\beta_2 = 0.99$, $\varepsilon = 1e\hat{}-8$) with the same learning rates of value 0.001 for the generator and discriminator and batch size of 64. Images in the training dataset are augmented by a large set of operations with probability 1, including geometrical transformations of horizontal flipping, isotropic and anisotropic scaling (scale_std = 0.2), arbitrary rotations, fractional translation (xfrac_std = 0.125), and color transformations in brightness (brightness_std = 0.2), contrast (contrast_std = 0.5), luma flip, hue rotation (hue_max = 1), and saturation (saturation_std = 1) [59]. The styleGAN2 model is trained for 16 epochs and generated images are of size 256 × 256 pixels.

B. Training Details: Latent Diffusion

Both the KL- and VQ-autoencoder are trained for 10 epochs with batch size 10 and a constant learning rate of $4.5*10^{-6}$. The DMs are trained with a constant learning rate of value 1e-6, batch size 64, and 30 numbers of epochs. During the fine-tuning step of the stable diffusion model, the autoencoder is frozen and only the diffusion process in latent space is trained. This fine-tuning is performed for a constant learning rate of 1e-6, batch size 16, and 20 epochs. For all LDMs, only random rotations by multiples of 90° are used to augment images.

C. Training Details: Multiclass Tissue Classifier

The classifiers are trained for 15 epochs, with batch size 64, and learning rate 1e-3 using the Adam optimizer [58] with a cyclic learning

rate scheduler as provided by fastai [60]. The models are evaluated on a held-out validation set from the training set after every epoch. The model with the lowest loss on the validation set is used to test it on the test dataset.

References

[1] O.L. Saldanha, C.M.L. Loeffler, J.M. Niehues, M. van Treeck, T.P. Seraphin, K. J. Hewitt, D. Cifci, G.P. Veldhuizen, S. Ramesh, A.T. Pearson, J.N. Kather, Self-supervised deep learning for pan-cancer mutation prediction from histopathology, bioRxiv (2022), https://doi.org/10.1101/2022.09.15.507455, 2022.09.15.507455.

[2] S.J. Wagner, D. Reisenbüchler, N.P. West, J.M. Niehues, J. Zhu, S. Foersch, G. P. Veldhuizen, P. Quirke, H.I. Grabsch, P.A. van den Brandt, G.G.A. Hutchins, S. D. Richman, T. Yuan, R. Langer, J.C.A. Jenniskens, K. Offermans, W. Mueller, R. Gray, S.B. Gruber, J.K. Greenson, G. Rennert, J.D. Bonner, D. Schmolze, J. Jonnagaddala, N.J. Hawkins, R.L. Ward, D. Morton, M. Seymour, L. Magill, M. Nowak, J. Hay, V.H. Koelzer, D.N. Church, TransSCOT consortium, C. Matek, C. Geppert, C. Peng, C. Zhi, X. Ouyang, J.A. James, M.B. Loughrey, M. Salto-Tellez, H. Brenner, M. Hoffmeister, D. Truhn, J.A. Schnabel, M. Boxberg, T. Peng, J. N. Kather, Transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study, Cancer Cell 41 (9) (2023 Sep 11) 1650–1661.e4, https://doi.org/10.1016/j.ccell.2023.08.002. Epub 2023 Aug 30. PMID: 37652006; PMCID: PMC10507381.

[3] A. Echle, N. Ghaffari Laleh, P. Quirke, H.I. Grabsch, H.S. Muti, O.L. Saldanha, S. F. Brockmoeller, P.A. van den Brandt, G.G.A. Hutchins, S.D. Richman, K. Horisberger, C. Galata, M.P. Ebert, M. Eckardt, M. Boutros, D. Horst, C. Reissfelder, E. Alwers, T.J. Brinker, R. Langer, J.C.A. Jenniskens, K. Offermans, W. Mueller, R. Gray, S.B. Gruber, J.K. Greenson, G. Rennert, J.D. Bonner, D. Schmolze, J. Chang-Claude, H. Brenner, C. Trautwein, P. Boor, D. Jaeger, N. T. Gaisa, M. Hoffmeister, N.P. West, J.N. Kather, Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a pre-screening tool for clinical application, ESMO Open 7 (2022) 100400.

[4] J.M. Niehues, P. Quirke, N.P. West, H.I. Grabsch, M. van Treeck, Y. Schirris, G. P. Veldhuizen, G.G.A. Hutchins, S.D. Richman, S. Foersch, T.J. Brinker, J. Fukuoka, A. Bychkov, W. Uegami, D. Truhn, H. Brenner, A. Brobeil, M. Hoffmeister, J. N. Kather, Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: a retrospective multi-centric study, Cell Rep Med (2023) 100980.

[5] Y. Schirris, E. Gavves, I. Nederlof, H.M. Horlings, J. Teuwen, DeepSMILE: contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer, Med. Image Anal. 79 (2022) 102464, https://doi.org/10.1016/j.media.2022.102464.

[6] D. Cifci, S. Foersch, J.N. Kather, Artificial intelligence to identify genetic alterations in conventional histopathology, J. Pathol. 257 (2022) 430–444.

[7] J.M. Dolezal, R. Wolk, H.M. Hieromnimon, F.M. Howard, A. Srisuwananukorn, D. Karpeyev, S. Ramesh, S. Kochanny, J.W. Kwon, M. Agni, R.C. Simon, C. Desai, R. Kherallah, T.D. Nguyen, J.J. Schulte, K. Cole, G. Khramtsova, M.C. Garassino, A. N. Husain, H. Li, R. Grossman, N.A. Cipriani, A.T. Pearson, Deep learning generates synthetic cancer histology for explainability and education, arXiv [eess.IV], http://arxiv.org/abs/2211.06522, 2022.

[8] M.A. Berbís, D.S. McClintock, A. Bychkov, J. Van der Laak, L. Pantanowitz, J. K. Lennerz, J.Y. Cheng, B. Delahunt, L. Egevad, C. Eloy, A.B. Farris, F. Fraggetta, R. García Del Moral, D.J. Hartman, M.D. Herrmann, E. Hollemans, K.A. Iczkowski, A. Karsan, M. Kriegsmann, M.E. Salama, J.H. Sinard, J.M. Tuthill, B. Williams, C. Casado-Sánchez, V. Sánchez-Turrión, A. Luna, J. Aneiros-Fernández, J. Shen, Computational pathology in 2030: a Delphi study forecasting the role of AI in pathology within the next decade, EBioMedicine 88 (2023) 104427.

[9] W.M. Grady, J.M. Carethers, Genomic and epigenetic instability in colorectal cancer pathogenesis, Gastroenterology 135 (2008) 1079–1099.

[10] R. Osuala, K. Kushibar, L. Garrucho, A. Linardos, Z. Szafranowska, S. Klein, B. Glocker, O. Diaz, K. Lekadir, Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging, Med. Image Anal. 84 (2023) 102704.

[11] M. Macenko, M. Niethammer, J.S. Marron, D. Borland, J.T. Woosley, X. Guan, C. Schmitt, N.E. Thomas, A method for normalizing histology slides for quantitative analysis, in: 2009 IEEE International Symposium on Biomedical Imaging: from Nano to Macro, 2009, pp. 1107–1110.

[12] S.J. Wagner, N. Khalili, R. Sharma, M. Boxberg, C. Marr, W. de Back, T. Peng, Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, 2021, pp. 257–266.

[13] A.B. Levine, J. Peng, D. Farnell, M. Nursey, Y. Wang, J.R. Naso, H. Ren, H. Farahani, C. Chen, D. Chiu, A. Talhouk, B. Sheffield, M. Riazy, P.P. Ip, C. Parra-Herran, A. Mills, N. Singh, B. Tessier-Cloutier, T. Salisbury, J. Lee, T. Salcudean, S. J.M. Jones, D.G. Huntsman, C.B. Gilks, S. Yip, A. Bashashati, Synthesis of diagnostic quality cancer pathology images by generative adversarial networks, J. Pathol. (2020), https://doi.org/10.1002/path.5509.

[14] J. Krause, H.I. Grabsch, M. Kloor, M. Jendrusch, A. Echle, R.D. Buelow, P. Boor, T. Luedde, T.J. Brinker, C. Trautwein, A.T. Pearson, P. Quirke, J. Jenniskens, K. Offermans, P.A. van den Brandt, J.N. Kather, Deep learning detects genetic alterations in cancer histology generated by adversarial networks, J. Pathol. 254 (2021) 70–79.

[15] R.J. Chen, M.Y. Lu, T.Y. Chen, D.F.K. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare, Nat. Biomed. Eng. 5 (2021) 493–497.

[16] K. Falahkheirkhah, T. Guo, M. Hwang, P. Tamboli, C.G. Wood, J.A. Karam, K. Sircar, R. Bhargava, A generative adversarial approach to facilitate archival-quality histopathologic diagnoses from frozen tissue sections, Lab. Invest. 102 (2022) 554–559.

[17] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, arXiv [stat.ML] (2014). http://arxiv.org/abs/1406.2661.

[18] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, arXiv [cs.LG] (2020) 6840–6851, in: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html. (Accessed 28 March 2023).

[19] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, PMLR, Lille, France, 2015, pp. 2256–2265.

[20] G. Müller-Franzes, J.M. Niehues, F. Khader, S.T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, S. Nebelung, J.N. Kather, D. Truhn, Diffusion probabilistic models beat GANs on medical images, arXiv [eess.IV], http://arxiv.org/abs/2212.07501, 2022.

[21] G. Müller-Franzes, J.M. Niehues, F. Khader, S.T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, J.N. Kather, D. Truhn, A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis, Sci. Rep. 13 (2023) 12098.

[22] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv [cs.LG] (2015). http://arxiv.org/abs/1511.06434.

[23] P. Dhariwal, A. Nichol, Diffusion models beat GANs on image synthesis, arXiv [cs.LG] (2021) 8780–8794, in: https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html. (Accessed 6 April 2023).

[24] J. Ho, C. Saharia, W. Chan, D.J. Fleet, M. Norouzi, T. Salimans, Cascaded diffusion models for high fidelity image generation, J. Mach. Learn. Res. 23 (2022) 1–33.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, arXiv [cs.CV] (2021). http://arxiv.org/abs/2112.10752.

[26] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with CLIP latents, arXiv [cs.CV], http://arxiv.org/abs/2204.06125, 2022.

[27] P. Chambon, C. Bluethgen, C.P. Langlotz, A. Chaudhari, Adapting pretrained vision-language foundational models to medical imaging domains, arXiv [cs.CV], http://arxiv.org/abs/2210.04133, 2022.

[28] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J.M. Z. Chaves, T.M. Abraham, S. Purohit, C.P. Langlotz, A. Chaudhari, RoentGen: vision-language foundation model for chest X-ray generation, arXiv [cs.CV], http://arxiv.org/abs/2211.12737, 2022.

[29] P.A. Moghadam, S. Van Dalen, K.C. Martin, J. Lennerz, S. Yip, H. Farahani, A. Bashashati, A Morphology Focused Diffusion Probabilistic Model for Synthesis of Histopathology Images, 2022, pp. 2000–2009, arXiv [eess.IV], https://openaccess.thecvf.com/content/WACV2023/html/Moghadam_A_Morphology_Focused_Diffusion_Probabilistic_Model_for_Synthesis_of_Histopathology_WACV_2023_paper.html. (Accessed 4 August 2023).

[30] M.U. Akbar, W. Wang, A. Eklund, Beware of diffusion models for synthesizing medical images – A comparison with GANs in terms of memorizing brain tumor images, arXiv [eess.IV], http://arxiv.org/abs/2305.07644, 2023.

[31] S.U.H. Dar, A. Ghanaat, J. Kahmann, I. Ayx, T. Papavassiliu, S.O. Schoenberg, S. Engelhardt, Investigating data memorization in 3D latent diffusion models for medical image synthesis, arXiv [cs.CV], http://arxiv.org/abs/2307.01148, 2023.

[32] J. Ye, H. Ni, P. Jin, S.X. Huang, Y. Xue, Synthetic augmentation with large-scale unconditional pre-training, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Springer Nature Switzerland, 2023, pp. 754–764.

[33] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, arXiv [cs.NE] (2018). http://arxiv.org/abs/1812.04948.

[34] J.N. Kather, N. Halama, A. Marx, 100,000 histological images of human colorectal cancer and healthy tissue. https://doi.org/10.5281/zenodo.1214456, 2018.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, arXiv [cs.LG] (2017). http://arxiv.org/abs/1706.08500.

[36] L. Gmail, G. Hinton, Visualizing Data using t-SNE. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl, 2008. (Accessed 4 August 2023).

[37] C.E. Shannon, A mathematical theory of communication, SIGMOBILE Mob. Comput. Commun. Rev. 5 (2001) 3–55.

[38] T.M. Cover, Elements of information theory, n.d. https://www.academia.edu/download/58191902/Elements_of_Information_Theory_Elements.pdf. (Accessed 29 April 2023).

[39] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, arXiv [cs.NE] (2017). http://arxiv.org/abs/1710.10196.

[40] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv [stat.ML] (2013). http://arxiv.org/abs/1312.6114v11.

[41] A. van den Oord, O. Vinyals, Koray Kavukcuoglu, Neural discrete representation learning, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017.

[42] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, A. Komatsuzaki, LAION-400M: open dataset of CLIP-filtered 400 million image-text pairs, arXiv [cs.CV], http://arxiv.org/abs/2111.02114, 2021.
[43] Y. Song, S. Ermon, Generative Modeling by Estimating Gradients of the Data Distribution, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc, 2019.
[44] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv [cs.LG] (2020). http://arxiv.org/abs/2010.02502.
[45] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, 2015, pp. 234–241.
[46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, arXiv [cs.CL] (2018). http://arxiv.org/abs/1810.04805.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017), in: https://proceedings.neurips.cc/paper/7181-attention-is-all.
[48] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers 2, 2003, pp. 1398–1402, 2003.
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, arXiv [cs.CV] (2015) 2818–2826. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html. (Accessed 30 March 2023).
[50] G. Parmar, R. Zhang, J.-Y. Zhu, On aliased resizing and surprising subtleties in GAN evaluation, arXiv [cs.CV], http://arxiv.org/abs/2104.11222, 2021.
[51] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2015, pp. 770–778, arXiv [cs.CV], http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html. (Accessed 30 March 2023).
[52] J. Howard, S. Gugger, Fastai: a layered api for deep learning, Information 11 (2020) 108.
[53] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (1971) 378–382.
[54] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved Techniques for Training GANs, 2016 arXiv [cs.LG], https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html. (Accessed 6 April 2023).
[55] L. Theis, A. van den Oord, M. Bethge, A note on the evaluation of generative models, arXiv [stat.ML] (2015). http://arxiv.org/abs/1511.01844.
[56] L. Mescheder, A. Geiger, S. Nowozin, Which training methods for GANs do actually converge? arXiv [cs.LG] (10–15 Jul (2018) 3481–3490, in: https://proceedings.mlr.press/v80/mescheder18a.html.
[57] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of StyleGAN, arXiv [cs.CV], http://arxiv.org/abs/1912.04958, 2019.
[58] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv [cs.LG] (2014). http://arxiv.org/abs/1412.6980.
[59] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila, Training generative adversarial networks with limited data, arXiv [cs.CV], http://arxiv.org/abs/2006.06676, 2020.
[60] L.N. Smith, Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 464–472.