This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Adapting Pre-trained Generative Model to Medical Image for Data Augmentation

Zhouhang Yuan[1,2], Zhengqing Fang[1,2], Zhengxing Huang[2], Fei Wu[2], Yu-Feng Yao[1], and Yingming Li[3(✉)]

[1] School of Medicine, Zhejiang University, Hangzhou, China
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[3] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
yingming@zju.edu.cn

Abstract. Deep learning-based medical image recognition requires a large number of expert-annotated data. As medical image data is often scarce and class imbalanced, many researchers have tried to synthesize medical images as training samples. However, the quality of the generated data determines the effectiveness of the method, which in turn is related to the amount of data available for training. To produce high-quality data augmentation in few-shot settings, we try to adapt large-scale pre-trained generative models to medical images. Specifically, we adapt MAGE (a masked image modeling-based generative model) as the pre-trained generative model, and then an Adapter is implemented within each layer to learn class-wise medical knowledge. In addition, to reduce the complexity caused by high-dimensional latent space, we introduce a vector quantization loss as a constraint during fine-tuning. The experiments are conducted on three different medical image datasets. The results show that our methods produce more realistic augmentation samples than existing generative models, with whom the classification accuracy increased by 5.16%, 2.74% and 3.62% on the three datasets respectively. The results demonstrate that adapting pre-trained generative models for medical image synthesis is a promising way in limited data situations.

Keywords: Medical Image Synthesis · Large Pre-trained Generative Model · Data Augmentation · Medical Image Classification.

1 Introduction

Deep learning networks trained on extensive medical image datasets have good representation and recognition abilities, making them useful in clinical diagnosis and classification [1]. However, large medical image datasets with high-quality labels are still rare due to the high annotation costs. Moreover, medical image datasets are typically collected during clinical diagnosis and treatment, often with class imbalance due to the sample scarcity, *e.g.* the number of normal samples is usually much higher than that of the diseased ones, or the number

of disease samples across categories may vary. The imbalance might lead to classification performance decrease [2].

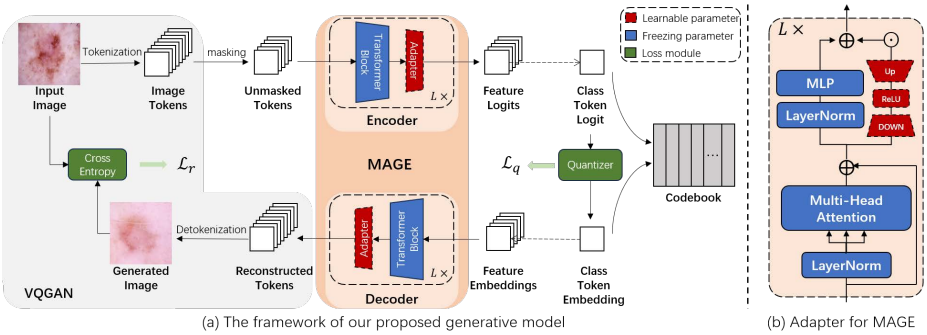
A potential solution is to obtain synthetic medical images as augmented training data via generative models. A generative model learns the distribution of the dataset and then generates synthetic data by sampling from the learned distribution. The representative generative techniques include Generative Adversarial Networks (GANs) [3], Variational Autoencoders (VAEs) [4], Flow-based models [5], and Diffusion-based models [6]. They have been extensively used in medical images for data augmentation, such as cervical cell pathology section images [7–9], fundus images [10], and chest X-ray images [11, 12]. However, it is important to note that models trained from scratch may be negatively impacted by insufficient data in a certain category within the dataset. This is because synthetic low-quality training data can harm the classifier's performance. Furthermore, a generative model trained from scratch only simulates the distribution of the given training data, meaning that all of its knowledge is derived from existing data, thus leading to limited performance improvement.

A novel way to improve the quality of synthetic data without adding additional data is to introduce a large pre-trained model. The large pre-trained generative model is obtained by pre-training on a large number of unsupervised web-scale images, so it has prior knowledge and is allowed to generalize to novel tasks with only a small number of samples. We proposed a large pre-trained generative model-based data augmentation method that can be applied to medical image datasets with few images. Specifically, a large pre-trained generative model based on masked image modeling (MAGE) [13] is chosen as the pre-trained model. The model is pre-trained on the ImageNet [14] dataset, surpassing Diffusion's performance on the same generation task while using fewer computing resources. To generate different class samples, an Adapter [15] layer is implemented on the Encoder and Decoder modules of MAGE, and trained on the image features represented by VQGAN [16]. Finally, two loss functions are used to constrain training. Along with the basic image reconstruction loss, we also introduce a vector quantization loss. This loss reduces the dimensionality of the feature latent space between the Encoder and Decoder models, making it easier to train with a small sample size and improving the generative model's performance.

The method was evaluated on three medical image datasets: Ham10000 [17], ODIR-5k [18], and Kera-3k. Experimental results show that using our method to generate samples as data augmentation can improve the classification performance on downstream tasks, and is superior to commonly used generative models such as StyleGAN2 [19], FastGAN [20] and Diffusion [6].

2 Method

Figure 1 (a) shows the overall structure of our approach. Our approach contains three steps. At first a well-trained VQGAN [16] encoder converts medical images from pixels to tokens. In the second step, we adapt the foundation generative



(a) The framework of our proposed generative model (b) Adapter for MAGE

Fig. 1. The framework of our method. We add an Adapter in each Transformer block and apply the quantization loss $\mathcal{L}_q$ to adapt for medical images.

model, MAGE, with parameter-efficient training which can encode medical images into features and generate images simultaneously. At last, we introduce vector quantization loss to reduce the difficulty of learning latent space in few-shot conditions.

2.1 Preliminary

The main structure of the MAGE model consists of a fixed VQGAN tokenizer, and a ViT-based encoder-decoder structure. VQGAN tokenizes the image into a sequence of semantic tokens $T = [t_i]_{i=1}^{N}, t_i \in \mathbb{R}^D$ of the image, where N is the token sequence length. Before encoding, we concatenate a learnable class token $[t_0]$ to the input sequence, then feed the token sequence into a Vision Transformer (ViT) [21] encoder-decoder structure, its formulaic expression is as follows

$$Z_l = \{[z_i]_{i=0}^{N}\}_l = B_l(Z_{l-1}), l \in \{0, 1, \ldots, L-1\}, \tag{1}$$

where $B_l(\cdot)$ represents the number l block of the encoder or decoder, L is the block number, Z_i represents the feature output by number i of the encoder or decoder, where $Z_0 = \{[z_i]_{i=0}^{N}\}_0$ equals to $[t_i]_{i=0}^{N}$. Each block contains an MHA(Multi-Head Attention) module and an MLP network, shown as follows:

$$Z_l^{'} = MHA(LN(Z_{l-1})) + Z_{l-1}, l \in \{0, 1, \ldots, L-1\}, \tag{2}$$

$$Z_l = MLP(LN(Z_l^{'})), \tag{3}$$

where $Z^{'}$ is the intermediate variable, $LN(\cdot)$ represents the LayerNorm module.

Let $Y = [y_i]_{i=1}^{N}$ denote the latent tokens obtained from the tokenizer. The reconstruction loss is a cross-entropy loss between the ground-truth one-hot tokens and the output of the decoder:

$$\mathcal{L}_{reconstructive} = -\mathbb{E}_{Y \in \mathcal{D}}(\sum \log\, p(y_i|Y)). \tag{4}$$

MAGE utilizes iterative decoding in MaskGIT [22] to fill in the mask token and generate the image iteratively. To generate an image during inference, it begins with a blank canvas with all tokens masked out.

2.2 Adapter for MAGE

Performing full fine-tuning on the MAGE model to achieve generation will cause a catastrophic forgetting problem [23], resulting in overwriting existing representations in the model. Therefore, we use the Adapter method to fine-tune the MAGE model. An Adapter is a neural network module that usually has a small number of parameters. By inserting the Adapter into the pre-trained model, one can keep the original parameters of the pre-trained model unchanged when fine-tuning on new datasets for efficient training and preventing catastrophic forgetting.

In our design, the Adapter method is introduced into each block B of both encoder and decoder. We demonstrate the details inside one block in Figure 1 (b). The red part is the Adapter, denoted as $A(\cdot)$, which is a bottleneck module that contains a down-projection layer with parameters $\mathbf{W}_{down}$ and an up-projection layer with parameters $\mathbf{W}_{up}$. The ReLU layer stands for a non-linear activation function.

$$A(Z_l^{'}) = ReLU(LN(Z'_l) \cdot \mathbf{W}_{down}) \cdot \mathbf{W}_{up}, \tag{5}$$

Where the parameters of $\mathbf{W}_{down}$ use Kaiming uniform initialize [24] and the parameters of $\mathbf{W}_{up}$ is initialized as zero. l represents the number of the block. $A(\cdot)$ is connected to the original MLP network (blue part) through the residual connection via a scale factor s. With Adapter, Equation 3 becomes:

$$Z_l = MLP(LN(Z^{'}{}_l)) + s \cdot A(Z_l^{'}), \tag{6}$$

During training, only the parameters of $A(\cdot)$ in Equation 6 are learnable.

2.3 Quantization loss

After encoding, the images become the feature vectors $Z_{L-1} \in \mathbb{R}^{(N+1)\times D}$, which are complex and contain redundant information. To compress these learned features, vector quantization (VQ) methods have been proposed to construct a dictionary of discrete vectors to approximate the actual continuous vectors [25].

The VQ method requires a latent embedding space $q \in \mathbb{R}^{K\times D}$ where K is the size of a discrete latent space. We take $z_{0,L-1}$ as the quantization module input, where $z_{0,L-1}$ is the class token of Z_{L-1} that contains the global information, as shown in Figure 1 (a). The VQ objective uses the l_2 error to move the embedding vectors q_i towards the feature logits $z_{0,L-1}$,

$$\mathcal{L}_{quantization} = ||sg\left[z_{0,L-1}\right] - q_i||_2^2 + \alpha||z_{0,L-1} - sg\left[q_i\right]||_2^2, \tag{7}$$

where $sg[\cdot]$ stands for the stop-gradient operator that is defined as an identity at forward computation time and has zero partial derivatives, thus effectively constraining its operand to be a no-updated constant.

The final loss function consists of the reconstruction loss and quantization loss, and a coefficient to control the loss ratio (default to 1), the calculation formula is as follows,

$$\mathcal{L} = \mathcal{L}_{reconstructive} + \beta \cdot \mathcal{L}_{quantization}. \tag{8}$$

3 Experiments

3.1 Setup

Dataset. The following three medical image datasets are employed in this study: 1)HAM10000 [17]: a dermatoscopic images dataset for skin lesions, 2)ODIR-5k [18]: a fundus images dataset for diabetic retinopathy classification, 3)Kera-3k: an anterior segment image for keratitis classification. The size of these datasets is shown in Table 1. We varied the quantity of generated data for different categories in the dataset to balance overall numbers with original training data. However, for categories with ample original data, like 'nv' in HAM10000 and 'N' in ODIR-5k, strict quantity adjustments might reduce augmentation performance. Thus, we adjusted generated data amounts separately for these categories.

Baselines. In limited data situations, some models like VAE may be hard to train or easily break down, so we chose some commonly used generation methods including FastGAN [20], StyleGAN2 [19], Latent Diffusion Model (LDM) [6] and MAGE [13] as our comparison methods. FastGAN and StyleGAN2 are two GAN-based generation methods that both contain a generator and a discriminator and train adversarial, both are trained from scratch. LDM is a model based on a diffusion algorithm and has a large parameter scale. In our study, we use the pre-trained LDM and finetune it with the Adapter named Lora. MAGE is a generative model based on masked image modeling (MIM), we use the model trained from the pre-trained MAGE model as a comparison.

Task and Metric (1) *Generation:* In this study, the Fréchet Inception Distance (FID) [26] was used to evaluate the generation quality by calculating the distance between the training sets and generation sets, with lower values indicating better quality. To avoid overfitting, we calculate the FID for both the training and test sets and report the results separately. (2) *Classification:* After generating medical images, we conduct experiments to evaluate the augmentation performance. The training set of the classification model is composed of the original training set and generation set from different generative models, which is class-balanced. The classifier is trained on this new training set and test on the original test set. The sample size of each data set is shown in Table 1. We fine-tune on two commonly used classification models, ViT [21] and Swin Transformer [27], and report two metric Accuracy (ACC) and Area Under the Curve (AUC) to measure classification performance.

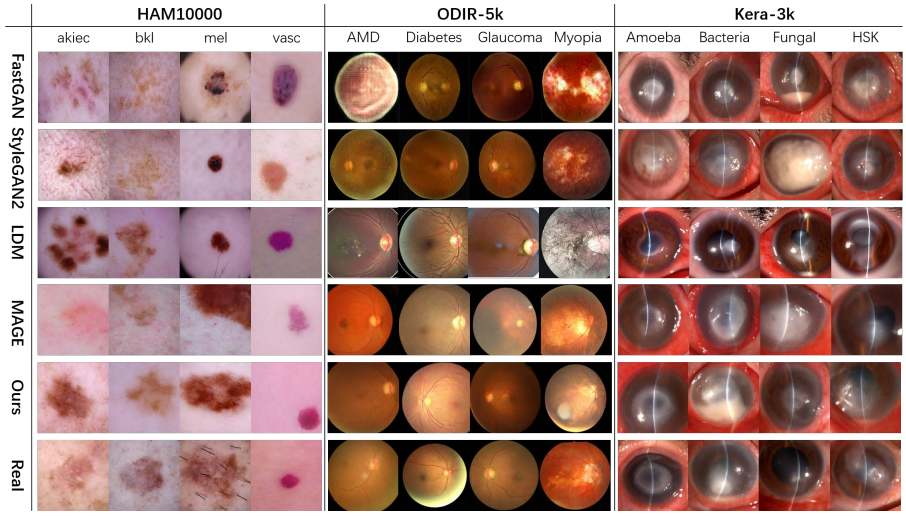


Fig. 2. Comparison of medical images generated by FastGAN, StyleGAN2, LDM, MAGE, and our method with real medical images, where four samples from different classes of each dataset are visualized. Each column represents the same category of images from different sources.

Implementation Details. We use the Adam [28] optimizer with a learning rate of 5×10^{-4} for training generative model and 1×10^{-3} for training classifier, where the generative model is trained for 1000 epochs and classifier is trained for 100 epochs.

3.2 Image generation performance comparison

In Figure 2, we randomly select and demonstrate medical images generated by several different generation methods, which can provide an intuitive reflection of the image generation quality of each method. From Figure 2, we can observe that the generated samples of FastGAN and StyleGAN2 are often blurred or distorted with poor image quality. Additionally, there are noticeable differences between the features of images generated by LDM and real images, particularly in the ODIR data set. The image generated by MAGE closely resembles the real image in style, but cannot learn correct disease characteristics, such as the inconsistent color and shape of lesions in skin diseases. Our method, based on MAGE, improves the generation of key features of disease lesions, resulting in skin lesions, blood vessels in fundus images, and keratitis lesions that are much closer to the real images.

After the sample analysis, we measured the FID metrics between the generation sets and the original datasets as a quantitative evaluation of the method generation performance. The FID results of the three datasets are recorded in Table 2, Table 3, Table 4, where we underline the optimal train set FID and the optimal test set FID in each column. Our approach has demonstrated op-

timal results in over half of the generation performance evaluations. On the HAM10000 and Kera-3k datasets, our method and FastGAN both demonstrate superior generation performance in certain categories. This is because the MAGE model, which serves as the foundation for our method, has weaker generation performance than FastGAN. However, our proposed method enhances the generation performance of the MAGE model, bringing it up to par with the optimal FastGAN method. On ODIR datasets, MAGE models demonstrate excellent generation performance, which reflects the advantage of higher adaptability of large pre-trained generative models. Our method further improves the generation performance based on the MAGE model and achieves the highest score in most categories.

Table 1. The size of three medical datasets.

Dataset	HAM10000						
	bkl	nv	df	mel	vasc	bcc	akiec
train	1099	6705	115	1113	142	514	327
test	217	908	44	171	35	93	43
generate	901	3295	1885	887	1858	1486	1673

Dataset	ODIR-5k							
	N	D	G	C	A	H	M	O
train	3104	1706	326	313	280	193	261	964
test	1255	708	128	142	123	80	98	533
generate	1896	1294	674	687	720	807	739	1036

Dataset	Kera-3k			
	amoeba	bacteria	fungal	hsk
train	294	437	818	1079
test	77	71	117	122
generate	706	563	1182	921

Table 2. Generation performance comparison of ODIR-5k using FID(↓).

Method	Dataset	Category							
		N	D	G	C	A	H	M	O
FastGAN	train	190.4	139.1	130.1	344.5	277.4	220.4	182.8	315.5
	test	191.1	141.4	136.5	356.2	280.5	233.8	196.3	324.6
StyleGAN2	train	132.4	157.4	115.6	98.0	142.7	140.5	117.9	51.0
	test	133.7	159.3	129.4	108.4	146.4	149.2	125.8	154.2
LDM	train	89.0	130.1	129.7	164.5	130.7	179.9	175.8	112.2
	test	90.3	132.1	136.0	187.5	142.0	196.5	194.0	120.4
MAGE	train	56.4	59.5	80.8	110.7	74.2	86.1	70.3	56.3
	test	59.5	61.6	97.7	134.3	80.9	93.0	89.2	57.7
Ours(w/o VQ)	train	46.8	55.5	69.4	96.8	70.5	85.0	76.8	63.1
	test	51.0	58.5	95.6	116.4	77.0	91.0	94.5	68.6
Ours	train	64.4	53.8	69.4	59.3	69.4	91.2	67.7	57.0
	test	69.3	56.0	95.6	83.6	79.4	101.8	87.8	59.19

Table 3. Generation performance comparison of HAM10000 using FID(↓).

Method	Dataset	Category						
		bkl	nv	df	mel	vasc	bcc	akiec
FastGAN	train	84.7	55.9	109.6	77.8	108.5	74.6	100.0
	test	114.9	67.0	140.2	114.3	189.0	116.9	160.6
StyleGAN2	train	99.4	103.3	150.1	112.4	167.9	92.3	134.7
	test	128.1	116.4	171.1	146.1	227.6	131.4	200.7
LDM	train	102.1	135.2	127.5	93.6	163.0	91.4	138.6
	test	124.1	145.8	158.0	124.1	231.2	130.5	206.7
MAGE	train	85.7	64.9	121.9	72.7	115.5	97.5	125.8
	test	121.4	77.7	161.8	114.2	194.8	142.0	197.7
Ours(w/o VQ)	train	99.6	77.7	136.1	85.3	126.2	100.8	133.7
	test	133.7	90.0	170.4	125.5	203.2	145.9	203.3
Ours	train	71.3	76.9	112.6	67.4	114.0	98.4	105.7
	test	114.5	91.5	161.1	112.7	184.5	141.7	166.6

Table 4. Generation performance comparison of Kera-3k using FID(↓).

Method	Dataset	Category			
		amoeba	bacteria	fungal	hsk
FastGAN	train	79.8	86.7	71.7	74.7
	test	123.3	119.1	115.0	116.7
StyleGAN2	train	83.1	86.0	77.3	73.1
	test	140.7	130.6	130.1	122.2
LDM	train	104.7	80.0	80.9	101.3
	test	164.0	126.7	118.2	137.9
MAGE	train	103.8	87.7	88.5	94.2
	test	161.0	155.0	134.2	144.1
Ours(w/o VQ)	train	83.9	75.4	64.7	66.9
	test	142.1	121.5	113.6	111.8
Ours	train	81.4	71.7	58.2	59.4
	test	136.6	118.0	107.6	105.5

Furthermore, we analyze the categories in which our method exhibits significant disadvantages. Notably, the three classes with the largest gaps are class N in the ODIR dataset (with an average difference of 18.0 from the optimal method), and classes nv and bcc in the HAM10000 dataset (with average differences of 22.8 and 24.3 from the optimal method, respectively). Table 1 shows that the first two categories have relatively high sample sizes of 3104 and 6705

Table 5. Performance comparison of different methods for data augmentation.

	Ham10000				ODIR-5k				Keratitis			
	ViT		Swin		ViT		Swin		ViT		Swin	
	ACC(%)	AUC	ACC(%)	AUC	ACC(%)	AUC	ACC(%)	AUC	ACC(%)	AUC	ACC(%)	AUC
Baseline	68.16	0.879	67.30	0.877	47.99	0.748	45.74	0.747	51.67	0.780	46.77	0.740
FastGAN	71.27	0.907	71.01	0.908	46.95	0.725	47.60	0.723	51.42	0.750	**52.45**	0.755
StyleGANv2	70.61	0.902	70.54	0.901	47.01	0.728	47.14	0.715	52.19	0.747	49.61	0.761
LDM	72.66	0.916	**73.19**	**0.927**	47.60	0.770	40.00	0.732	48.83	0.700	51.67	0.704
MAGE	72.86	0.910	69.62	0.897	49.33	0.781	48.77	0.766	51.67	0.764	50.90	0.770
Ours(w/o VQ)	72.07	0.912	71.34	0.918	**50.86**	0.782	47.89	0.751	50.12	0.729	49.61	0.726
Ours	**73.19**	**0.923**	72.60	0.921	50.30	**0.783**	**48.90**	**0.771**	**53.48**	**0.781**	52.19	**0.784**

respectively. This also proves that our method is more applicable in the case of limited sample size, while has no obvious performance advantage in the case of sufficient sample size.

3.3 Classification performance after augmentation

The way to use the generated data as data augmentation is to add it to the training set, which addresses the issue of limited data size and unbalanced category numbers in the original set, ultimately improving classification performance. Table 5 shows that our approach has yielded the best results in most cases (**marked bold**). Only on the HAM10000 dataset, when Swin Transformer is used as the classifier, the augmentation effect of LDM exceed that of our method. It is worth noting that the difference between the two is not large (only 0.59% on ACC and 0.006 on AUC), and in all other cases, the augmentation performance of our method is significantly better than that of LDM. In addition, we observed that in ODIR-5k dataset, when ViT was used as the classification model, the ablation method without VQ loss achieved better comprehensive classification performance, which was consistent with the data in Table 2, where the generation performance of ablation methods is better in some categories. The ODIR dataset has a large sample size, which may explain the performance loss of the VQ loss. This confirms that our method is more suitable for cases with small sample sizes. Overall, our method achieves optimal results in most indicators and suboptimal results in others. Additionally, our method outperforms all other methods in average classification performance, indicating that it has the best augmentation performance. Data augmentation using our method can significantly improve the performance of medical image classification tasks, among which the average ACC and AUC on HAM10000, ODIR-5k and Kera-3k datasets increased by 5.16%, 2.74%, 3.62%, and 0.042, 0.030, 0.023, respectively.

4 Discussion

In this study, we propose a data augmentation technique for medical images by adapting a large pre-trained genetive model. We introduce the Adapter and a

vector quantization loss to finetune MAGE, which is a generative data augmentation method that is more suitable for limited data size. Compared with the existing methods, our approach produces images that exhibit advantages in both visual sample evaluation and comprehensive quantitative evaluation. Additionally, using the generated samples led to a significant improvement in classification performance, as shown in Table 5. The limitation of our study is that we only use optical medical images for the experiment, which is closer to the general images. Different Adapter placement positions may be suitable for other modalities of medical images like X-ray or MRI, which can be an interesting topic for future works. In conclusion, our study shows that utilizing large pre-trained generative models as data augmentation sources is a promising approach for numerous medical image datasets with limited samples.

Acknowledgments. This study was funded by the National Key R&D Program of China (2023YFE0204200), the National Natural Science Foundation of China (Grant No. U20A20387), the Zhejiang Provincial Key Research and Development Project (Grant No. 2023C01043).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
2. Matloob Khushi, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A Hameed, Shahadat Uddin, Suhuai Luo, Xiaoyan Yang, and Maranatha Consuelo Reyes. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9:109960–109975, 2021.
3. Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.
4. Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.
5. Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 2021.
6. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
7. Zhenrong Shen, Maosong Cao, Sheng Wang, Lichi Zhang, and Qian Wang. Cellgan: Conditional cervical cell synthesis for augmenting cytopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 487–496. Springer, 2023.

8. Jiarong Ye, Haomiao Ni, Peng Jin, Sharon X Huang, and Yuan Xue. Synthetic augmentation with large-scale unconditional pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 754–764. Springer, 2023.
9. Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2023.
10. Zhihao Zhao, Junjie Yang, Shahrooz Faghihroohi, Kai Huang, Mathias Maier, Nassir Navab, and M Ali Nasseri. Label-preserving data augmentation in latent space for diabetic retinopathy recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 284–294. Springer, 2023.
11. Zeyi Hou, Ruixin Yan, Qizheng Wang, Ning Lang, and Xiuzhuang Zhou. Diversity-preserving chest radiographs generation from reports in one stage. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 482–492. Springer, 2023.
12. Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.
13. Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152, 2023.
14. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
15. Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.
16. Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
17. Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.
18. Ning Li, Tao Li, Chunyu Hu, Kai Wang, and Hong Kang. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In *Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Bench 2020, Virtual Event, November 15–16, 2020, Revised Selected Papers 3*, pages 177–193. Springer, 2021.
19. Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.
20. Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2020.
21. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg

Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
22. Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.
23. Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020.
24. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
25. Kuilin Chen and Chi-Guhn Lee. Incremental few-shot learning via vector quantization in deep embedded space. In *International Conference on Learning Representations*, 2020.
26. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
27. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
28. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.