Synthesis of Color Fundus Images Using an Autoregressive Model

Authors: Yiwei Chen, Yiwei Chen

Date: 2026-01-09T22:33:37+00:00

Abstract

Autoregressive models have achieved notable success in image generation, owing to their strong sequential modeling capabilities, flexible generative control, and excellent scalability across diverse visual tasks. Motivated by these advantages, we adopt an autoregressive approach to synthesize color fundus images. In particular, our method generates color fundus photographs for both healthy individuals and patients with moderate to severe non-proliferative diabetic retinopathy.

To assess the quality of the synthesized images, we compute the Fréchet Inception Distance (FID), obtaining scores of 40.14 for healthy fundus photographs and 41.21 for pathological fundus photographs. These results indicate that our approach outperforms a widely used diffusion-based model.

Full Text

Preamble

Synthesizing Color Fundus Images Using Autoregressive Models

Yiwei Chen*
Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China
yiwei.chen@sibet.ac.cn

Autoregressive models have demonstrated considerable success in image generation, attributed to their robust sequential modeling capabilities, flexible generative control, and strong scalability across diverse visual tasks. Motivated by these advantages, we adapted this approach to synthesize color fundus photographs. Specifically, our algorithm generates color fundus images of both healthy subjects and patients diagnosed with moderate to severe non-proliferative diabetic retinopathy. To evaluate the quality of the generated images, we computed the Fréchet Inception Distance score, achieving 40.14

for healthy fundus photographs and 41.21 for pathological fundus photographs. These results demonstrate superior performance compared to a widely-used diffusion model.

Keywords: image synthesis, autoregressive model, color fundus image, nonproliferative diabetic retinopathy

1 Introduction

The synthesis of high-quality medical images has emerged as a critical research direction in computational ophthalmology, addressing fundamental challenges in data scarcity, privacy preservation, and algorithmic development [**?**, **?**]. Color fundus photography, as a non-invasive and widely accessible diagnostic modality, serves as the cornerstone for screening and monitoring various retinal pathologies, particularly diabetic retinopathy (DR) [**?**, **?**].

However, the acquisition of large-scale, well-annotated fundus image datasets remains constrained by multiple factors, including patient privacy concerns, the requirement for expert annotation, and the inherent imbalance in disease prevalence [**?**, **?**]. These limitations have motivated substantial interest in generative modeling approaches capable of producing realistic synthetic fundus photographs.

Recent advances in deep generative models have demonstrated remarkable capabilities in natural image synthesis, with diffusion models emerging as the dominant paradigm [**?**, **?**]. These models have achieved state-of-the-art performance across various benchmarks by iteratively denoising random noise into coherent images. In the medical imaging domain, diffusion-based approaches have been explored for fundus photograph generation, showing promising results in capturing retinal structures [**?**]. Despite their success, diffusion models present certain practical limitations, including substantial computational requirements during inference, complex training dynamics, and the absence of explicit sequential modeling that could facilitate fine-grained control over image generation [**?**, **?**].

Concurrently, autoregressive (AR) models have experienced a renaissance in visual generation, driven largely by their unprecedented success in large language models (LLMs) [**?**, **?**]. The fundamental "next-token prediction" paradigm that powers LLMs has been successfully adapted to image synthesis through discrete visual tokenization [**?**, **?**]. Recent work has demonstrated that vanilla autoregressive models, without vision-specific inductive biases, can achieve competitive or superior performance compared to diffusion models when properly scaled [**?**]. These models offer several compelling advantages: they leverage the mature ecosystem of LLM optimization techniques, provide explicit sequential generation control, and demonstrate favorable scaling properties [**?**, **?**].

Building upon these developments, we investigate the application of autoregressive models for synthesizing color fundus photographs. Our approach utilizes the LlamaGen architecture [**?**], which has shown strong performance on natural

images, to generate color fundus images. We specifically target the generation of both healthy fundus photographs and images exhibiting moderate to severe non-proliferative diabetic retinopathy (NPDR), representing clinically significant pathological variations. Through systematic evaluation using the Fréchet Inception Distance (FID) metric, our method achieves scores of 40.14 for healthy images and 41.21 for pathological images, demonstrating substantial improvements over established diffusion-based approaches [**?**].

The contributions of this work are threefold. First, we demonstrate that autoregressive models can effectively synthesize high-quality fundus photographs without requiring domain-specific architectural modifications. Second, we show that AR-based generation can capture both normal retinal anatomy and pathological features associated with diabetic retinopathy with superior fidelity compared to diffusion models. Third, we provide empirical evidence that computational efficiency and controllability advantages of AR models translate effectively to the medical imaging context. These findings suggest that autoregressive approaches represent a promising direction for medical image synthesis, potentially bridging the gap between advances in natural language processing and clinical image generation.

2.1 Generative Models for Medical Image Synthesis

The application of generative models to medical imaging has evolved significantly over the past decade. Early approaches primarily relied on Generative Adversarial Networks (GANs) [**?**, **?**], which demonstrated the capability to synthesize realistic medical images across various modalities. In ophthalmology, GAN-based methods have been employed to generate fundus photographs for data augmentation [**?**] and to simulate disease progression [**?**]. However, GANs are notorious for training instability, mode collapse, and the challenge of balancing the discriminator-generator dynamics [**?**].

The emergence of diffusion models marked a paradigm shift in image generation [**?**, **?**]. Denoising Diffusion Probabilistic Models (DDPMs) [**?**] and their variants have achieved state-of-the-art results in natural image synthesis by learning to reverse a gradual noising process. The improved DDPM (iDDPM) [**?**] introduced learned variance schedules and hybrid training objectives, further enhancing image quality and log-likelihood. In medical imaging, diffusion models have shown promise for various tasks including image reconstruction [**?**], super-resolution [**?**], and synthesis [**?**].

Specifically for fundus photography, recent work has demonstrated that DDPMs can generate synthetic retinal images that capture both anatomical structures and pathological features [**?**]. Despite these achievements, diffusion models typically require hundreds of denoising steps during inference, leading to substantial computational costs that may limit their practical deployment in clinical settings [**?**, **?**].

2.2 Autoregressive Models for Visual Generation

Autoregressive models have a long history in sequence modeling, achieving remarkable success in natural language processing [**?**, **?**]. The core principle of predicting the next token given previous context has proven to be a powerful and scalable approach. The adaptation of this paradigm to visual generation began with Vector Quantized Variational AutoEncoders (VQ-VAE) [**?**], which discretize continuous image pixels into learnable codebook entries. Building upon this foundation, models such as VQGAN [**?**] improved image quality through adversarial training and perceptual losses.

The integration of transformer architectures [**?**] with discrete visual representations led to significant advances. DALL-E [**?**] demonstrated that autoregressive transformers could generate diverse images from textual descriptions by treating both text and image tokens uniformly. Subsequent work including Parti [**?**] and Make-A-Scene [**?**] further explored the scalability and controllability of AR-based image generation. However, these early models often incorporated vision-specific inductive biases and architectural modifications, making it unclear whether vanilla autoregressive approaches could achieve competitive performance.

Recent developments have challenged this assumption. LlamaGen [**?**] demonstrated that applying the original "next-token prediction" paradigm of large language models directly to visual generation, without specialized architectural adaptations, can achieve state-of-the-art performance when properly scaled. This work showed that vanilla AR models outperformed established diffusion models like LDM [**?**] and DiT [**?**] on ImageNet benchmarks. The key enablers included improved image tokenizers with lower reconstruction error, better scalability properties through architecture choices borrowed from LLMs [**?**, **?**], and optimized inference through LLM serving frameworks [**?**].

Alternative approaches to AR-based visual generation have emerged. Masked generative models such as MaskGIT [**?**] and MAGVIT [**?**] employ bidirectional attention and iterative decoding, deviating from the strictly unidirectional generation of traditional AR models. The recently proposed Visual Autoregressive modeling (VAR) [**?**] introduces a coarse-to-fine "next-scale prediction" paradigm, achieving superior performance on ImageNet generation. However, these methods incorporate vision-specific modifications that distinguish them from the pure next-token prediction approach.

2.3 Fundus Photograph Synthesis with Diffusion Models

The application of diffusion models to fundus photograph synthesis has been explored in recent work [**?**]. This study investigated the feasibility of using DDPMs for generating fundus images with a relatively small dataset of 1,000 healthy retinal images. The authors trained models at 128×128 resolution and compared performance against Progressive Growing GAN (PGGAN) [**?**]. While DDPMs successfully generated synthetic fundus photographs without mode col-

lapse, the study revealed several limitations when working with limited data: the models required extensive training time (approximately 250 hours for 128×128 images), produced relatively blurred anatomical features compared to GANs, and achieved inferior FID scores (65.605) compared to PGGAN (41.761). The authors concluded that diffusion models need larger datasets and computational resources to achieve competitive performance in domain-specific medical imaging tasks.

These findings highlight a critical gap: while diffusion models have shown strong performance on large-scale natural image datasets, their effectiveness on smaller, domain-specific medical imaging datasets remains limited. This observation motivates the exploration of alternative generative approaches that may offer better sample efficiency and computational tractability for medical image synthesis.

2.4 Synthesis of Diabetic Retinopathy Images

Generating synthetic images of pathological conditions presents additional challenges beyond normal anatomy. For diabetic retinopathy, the key pathological features include microaneurysms, hemorrhages, exudates, and cotton-wool spots, which vary in size, location, and severity [**?**, **?**]. Previous work has explored GAN-based approaches for synthesizing DR images [**?**, **?**], but achieving realistic lesion characteristics while maintaining overall image coherence remains challenging. The ability to generate high-quality pathological images has significant implications for developing robust diagnostic algorithms, particularly for rare or severe disease stages where real data is scarce [**?**, **?**].

3.1 Overview

Our approach for synthesizing color fundus photographs follows a two-stage pipeline that has proven effective in autoregressive image generation [**?**]. The first stage involves training an image tokenizer that converts continuous pixel representations into discrete visual tokens. The second stage trains an autoregressive transformer model to generate these tokens conditioned on class labels. Figure 1 [Figure 1: see original paper] illustrates the overall architecture of our method.

3.2 Dataset

We utilized the OIA-DDR dataset [**?**], a publicly available collection of color fundus photographs annotated for diabetic retinopathy severity. From this dataset, we constructed a binary classification scheme suitable for conditional image generation. The first class comprised all healthy fundus images, totaling 6,266 samples. The second class combined moderate and severe non-proliferative diabetic retinopathy (NPDR) cases, yielding 4,713 samples. This grouping strategy was motivated by the clinical similarity between moderate and severe NPDR stages, both representing significant pathological progression that warrants close monitoring [**?**, **?**].

The dataset was partitioned following standard practices in machine learning evaluation. Approximately 50% of the data was allocated to training (3,133 healthy images and 2,356 pathological images), 20% to validation (1,253 healthy and 942 pathological), and 30% to testing (1,880 healthy and 1,415 pathological). Table 1 summarizes the dataset composition.

3.3 Image Preprocessing

Fundus photographs exhibit considerable variation in aspect ratio and field of view depending on the imaging device and acquisition protocol [**?**]. To ensure consistent input dimensions while preserving the circular region of interest containing retinal structures, we applied a center-cropping strategy. Specifically, each image was first cropped to extract the maximum inscribed square from its center, eliminating peripheral black borders that typically surround fundus images. The cropped images were subsequently resized to 256×256 pixels using bilinear interpolation. This resolution represents a practical compromise between computational efficiency and the preservation of clinically relevant details such as microaneurysms and small hemorrhages [**?**, **?**]. All images retained three color channels (RGB), as color information is diagnostically significant in fundus photography for identifying features such as exudates and hemorrhages [**?**].

3.4 Image Tokenization

The image tokenizer serves as the critical bridge between continuous pixel space and the discrete token space required by autoregressive models [**?**, **?**]. We employed a vector-quantized autoencoder architecture following the design principles established in recent work [**?**].

3.4.1 Architecture

The tokenizer consists of an encoder network, a vector quantization module, and a decoder network. The encoder transforms input images into a spatial feature map through a series of convolutional layers with downsampling operations. We utilized a downsampling ratio of 8, meaning that a 256×256 input image produces a 32×32 feature map. Each spatial position in this feature map is then mapped to its nearest neighbor in a learned codebook, yielding a grid of 1,024 discrete tokens per image.

The codebook was configured with 16,384 entries, each represented by an 8-dimensional embedding vector. This configuration follows recommendations from prior studies demonstrating that lower-dimensional codebook vectors combined with larger codebook sizes yield superior reconstruction quality and codebook utilization [**?**, **?**]. The ℓ_2-normalization was applied to codebook vectors to stabilize training and improve convergence [**?**].

3.4.2 Training Objective

The tokenizer was trained using a combination of reconstruction and adversarial losses. The reconstruction objective comprised a pixel-wise ℓ_2 loss and a perceptual loss computed using LPIPS [**?**], which measures feature-level similarity in a pretrained VGG network. An adversarial loss from a PatchGAN discriminator [**?**] was incorporated to enhance the sharpness and realism of reconstructed images. The commitment loss [**?**] was included to encourage encoder outputs to remain close to their assigned codebook entries, with a weighting factor of 0.25. The adversarial loss was weighted at 0.5 and activated after 20,000 training iterations to allow the autoencoder to first learn basic reconstruction before refining perceptual quality.

We utilized a pretrained tokenizer model that had been trained on ImageNet [**?**], leveraging transfer learning to adapt to the fundus image domain. This approach is justified by the observation that low-level visual features learned from natural images often transfer well to medical imaging tasks [**?**, **?**].

3.5 Autoregressive Generation Model

3.5.1 Architecture

The autoregressive generation model follows the Llama architecture [**?**], which has demonstrated strong scalability in language modeling tasks. This design choice enables direct adoption of optimization techniques developed for large language models without requiring vision-specific modifications. The model employs several architectural components that have proven effective in transformer-based language models.

Pre-normalization using RMSNorm [**?**] is applied before each attention and feed-forward layer, providing training stability. The SwiGLU activation function [**?**] replaces the standard GELU or ReLU in feed-forward networks, offering improved gradient flow. Rotary positional embeddings (RoPE) [**?**] encode position information directly into the attention mechanism. For image generation, we extended RoPE to two dimensions to capture the spatial structure of image tokens, following established practices [**?**, **?**].

We employed the GPT-XL configuration, comprising 36 transformer layers, a hidden dimension of 1,280, and 20 attention heads, yielding approximately 775 million parameters. This model size was selected based on the dataset scale and available computational resources, following scaling law observations suggesting that model capacity should be matched to training data volume [**?**, **?**].

3.5.2 Conditional Generation

Class-conditional generation was implemented by prepending a learnable class embedding to the sequence of image tokens [**?**, **?**]. The model learns to generate image tokens autoregressively, with each token prediction conditioned on the

class label and all previously generated tokens. To enable classifier-free guidance [**?**] during inference, we randomly replaced the class condition with a null embedding during training with probability 0.1. This technique allows trading off sample diversity for fidelity at generation time by interpolating between conditional and unconditional predictions.

3.5.3 Training Procedure

Prior to training the autoregressive model, we extracted discrete codes for all training images using the pretrained tokenizer. This preprocessing step significantly accelerates training by eliminating the need for repeated forward passes through the tokenizer during each epoch [**?**]. The extracted codes were stored as NumPy arrays along with their corresponding class labels.

The model was trained using the AdamW optimizer [**?**] with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.05. Gradient clipping with a maximum norm of 1.0 was applied to prevent training instability. The learning rate was set to 1×10^{-4}. We employed mixed-precision training using bfloat16 format to reduce memory consumption and accelerate computation on modern GPU architectures [**?**]. Dropout regularization was applied at multiple points in the architecture: token embedding dropout (0.1), attention dropout (0.1), and feed-forward network dropout (0.1). These regularization strategies help prevent overfitting, which is particularly important given the relatively modest size of our training dataset compared to natural image benchmarks.

Training was conducted using distributed data parallelism across multiple GPUs. For models exceeding single-GPU memory capacity, we employed Fully Sharded Data Parallelism (FSDP) [**?**], which distributes model parameters, gradients, and optimizer states across devices. The global batch size was set to 32, distributed across available GPUs. Training proceeded for 300 epochs, with checkpoints saved at regular intervals for evaluation.

3.5.4 Inference

During inference, image tokens were generated autoregressively starting from the class embedding token. At each step, the model predicts a probability distribution over the codebook, and the next token is sampled from this distribution. We employed nucleus sampling (top-p) [**?**] with $p = 1.0$ and temperature $= 1.0$ as the default configuration, following recommendations for maintaining sample diversity [**?**].

Classifier-free guidance was applied during inference to enhance image quality. The guided logit was computed as a weighted combination of conditional and unconditional predictions, with higher guidance scales producing sharper but potentially less diverse samples. We experimented with guidance scales ranging from 1.0 to 4.0. The generated token sequences were decoded through the tokenizer decoder to produce final RGB images at 256×256 resolution.

3.6 Evaluation Metrics

We adopted the FID [**?**] as the primary metric for assessing the quality of generated fundus images. FID measures the distance between the feature distributions of real and generated images, computed using activations from a pretrained Inception network. Lower FID values indicate greater similarity between generated and real image distributions. This metric has been widely adopted in generative modeling and provides a reliable assessment of both image quality and diversity [**?**, **?**].

FID was computed separately for healthy and pathological image classes to assess class-specific generation quality. For each class, we generated a number of synthetic images equal to the test set size and computed FID against the corresponding test set images. All FID calculations followed the standard protocol for consistency with prior work [**?**].

3.7 Implementation Details

All experiments were conducted using PyTorch [**?**] on NVIDIA A100 GPUs. The tokenizer utilized the VQ-GAN architecture with 8× downsampling, producing 32×32 token grids from 256×256 input images. Model compilation was enabled using torch.compile to optimize runtime performance. Training logs and checkpoints were managed systematically to ensure reproducibility.

Table 2 : Summary of key hyperparameters and architectural configurations.

Component	Configuration
Image resolution	256×256
Tokenizer downsampling	8×
Tokens per image	1024 (32×32)
Codebook size	16,384
Codebook dimension	8
Model architecture	GPT-XL
Transformer layers	36
Hidden dimension	1,280
Attention heads	20
Learning rate	1×10^{-4}
Weight decay	0.05
Batch size	32
Training epochs	300
Dropout rate	0.1
CFG dropout	0.1

4 Results

4.1 Quantitative Evaluation

We evaluated the performance of our autoregressive approach against the improved Denoising Diffusion Probabilistic Model (iDDPM) [**?**] using FID as the primary metric. Table 3 summarizes the quantitative results for both healthy and pathological fundus image generation.

Table 3: FID scores comparing the proposed autoregressive method with iDDPM for fundus image synthesis. Lower values indicate better performance.

Category	Our method (AR) [**?**]	iDDPM [**?**]	Improvement
Healthy fundus images	40.14	78.20	48.7%
Pathological fundus images (Moderate-Severe NPDR)	41.21	74.19	44.5%

Our autoregressive model achieved substantially lower FID scores across both image categories. For healthy fundus photographs, the proposed method attained an FID of 40.14 compared to 78.20 for iDDPM, representing a relative improvement of 48.7%. Similarly, for pathological images depicting moderate to severe NPDR, our approach achieved an FID of 41.21 versus 74.19 for the diffusion-based baseline, corresponding to a 44.5% reduction in FID.

These results demonstrate that the autoregressive paradigm provides meaningful advantages over diffusion models in this medical imaging context. The performance gap was consistent across both healthy and pathological categories, suggesting that the benefits of autoregressive generation extend to images with complex pathological features such as microaneurysms, hemorrhages, and exudates characteristic of diabetic retinopathy [**?**, **?**].

4.2 Qualitative Evaluation

Visual inspection of the generated samples revealed appreciable differences between the two approaches. Images synthesized by our autoregressive model generally exhibited sharper appearance and more coherent structural details compared to those produced by iDDPM. The retinal vasculature, optic disc boundaries, and overall image contrast appeared more clearly defined in the autoregressive outputs. For pathological images, the lesion features associated with diabetic retinopathy were rendered with better clarity by our method. In contrast, the diffusion model tended to produce outputs with somewhat blurred details and reduced sharpness across anatomical structures. These qualitative observations are consistent with the quantitative FID improvements reported

above, suggesting that the autoregressive paradigm offers advantages in preserving fine-grained visual details during fundus image synthesis.

Figure 2 [Figure 2: see original paper]: Visual comparison of synthesized fundus photographs. Top row: healthy fundus images. Bottom row: pathological fundus images with moderate to severe NPDR. Left panels in each row show results from the proposed autoregressive method; right panels show results from iDDPM.

5 Conclusions

This study investigated the application of autoregressive models for synthesizing color fundus photographs, demonstrating their effectiveness in generating both healthy retinal images and pathological images exhibiting moderate to severe non-proliferative diabetic retinopathy. Our approach, built upon the LlamaGen architecture, achieved FID scores of 40.14 for healthy fundus images and 41.21 for pathological images, representing improvements of 48.7% and 44.5% respectively compared to the improved DDPM baseline.

The results suggest several practical implications. The autoregressive paradigm offers a viable alternative to diffusion models for medical image synthesis, particularly when computational efficiency during inference is a consideration. The ability to generate pathological images with comparable quality to healthy images indicates that the sequential token prediction mechanism can capture disease-specific features such as microaneurysms and hemorrhages without requiring specialized architectural modifications.

References

[1] Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence. 2020;2(6):305-311.

[2] Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering. 2021;5(6):493-497.

[3] Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Medicine. 2018;1(1):39.

[4] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-2410.

[5] Wilkinson CP, Ferris FL, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110(9):1677-1682.

[6] Coyner AS, Swan R, Campbell JP, et al. Deep learning for image quality assessment of fundus images in retinopathy of prematurity. AMIA Annual Symposium Proceedings. 2018;2018:1224-1232.

[7] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems. 2020;33:6840-6851.

[8] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems. 2021;34:8780-8794.

[9] Kim HK, Ryu IH, Choi JY, Yoo TK. A feasibility study on the adoption of a generative denoising diffusion model for the synthesis of fundus photographs using a small dataset. Discover Applied Sciences. 2024;6:188.

[10] Song J, Meng C, Ermon S. Denoising diffusion implicit models. International Conference on Learning Representations. 2021.

[11] Lu C, Zhou Y, Bao F, Chen J, Li C, Zhu J. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems. 2022;35:5775-5787.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30:5998-6008.

[13] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020;33:1877-1901.

[14] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. International Conference on Machine Learning. 2021:8821-8831.

[15] Yu J, Xu Y, Koh JY, et al. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789. 2022.

[16] Sun P, Jiang Y, Chen S, et al. Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525. 2024.

[17] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. 2023.

[18] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. 2023.

[19] Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM. 2020;63(11):139-144.

[20] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: An overview. IEEE Signal Processing Magazine. 2018;35(1):53-65.

[21] Zhao H, Li H, Maurer-Stroh S, Cheng L. Synthesizing retinal and neuronal images with generative adversarial nets. Medical Image Analysis. 2018;49:14-26.

[22] Costa P, Galdran A, Meyer MI, et al. End-to-end adversarial retinal image synthesis. IEEE Transactions on Medical Imaging. 2018;37(3):781-791.

[23] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. International Conference on Machine Learning. 2017:214-223.

[24] Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. International Conference on Machine Learning. 2015:2256-2265.

[25] Nichol AQ, Dhariwal P. Improved denoising diffusion probabilistic models. International Conference on Machine Learning. 2021:8162-8171.

[26] Chung H, Sim B, Ryu D, Ye JC. Improving diffusion models for inverse problems using manifold constraints. Advances in Neural Information Processing Systems. 2022;35:25683-25696.

[27] Saharia C, Ho J, Chan W, et al. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;45(4):4713-4726.

[28] Kazerouni A, Aghdam EK, Heidari M, et al. Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis. 2023;88:102846.

[29] Kwon M, Jeong J, Uh Y. Diffusion models already have a semantic latent space. International Conference on Learning Representations. 2023.

[30] Chen T, Zhang R, Hinton G. Analog bits: Generating discrete data using diffusion models with self-conditioning. International Conference on Learning Representations. 2023.

[31] Van Den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. Advances in Neural Information Processing Systems. 2017;30:6306-6315.

[32] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:12873-12883.

[33] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations. 2021.

[34] Gafni O, Polyak A, Ashual O, et al. Make-a-scene: Scene-based text-to-image generation with human priors. European Conference on Computer Vision. 2022:89-106.

[35] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:10684-10695.

[36] Peebles W, Xie S. Scalable diffusion models with transformers. IEEE/CVF International Conference on Computer Vision. 2023:4195-4205.

[37] Su J, Lu Y, Pan S, Murtadha A, Wen B, Liu Y. RoFormer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864.

[38] Shazeer N. GLU variants improve transformer. arXiv preprint arXiv:2002.05202. 2020.

[39] Kwon W, Li Z, Zhuang S, et al. Efficient memory management for large language model serving with PagedAttention. ACM SIGOPS Operating Systems Review. 2023;57(1):611-626.

[40] Chang H, Zhang H, Jiang L, Liu C, Freeman WT. MaskGIT: Masked generative image transformer. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:11315-11325.

[41] Yu L, Shi Y, Wang Y, et al. MAGVIT: Masked generative video transformer. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:10459-10469.

[42] Tian K, Jiang Y, Yuan Z, Peng B, Wang L. Visual autoregressive modeling: Scalable image generation via next-scale prediction. arXiv preprint arXiv:2404.02905. 2024.

[43] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. International Conference on Learning Representations. 2018.

[44] Yau JW, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012;35(3):556-564.

[45] Wong TY, Cheung CM, Larsen M, Sharma S, Simó R. Diabetic retinopathy. Nature Reviews Disease Primers. 2016;2(1):16012.

[46] Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmology. 2017;135(11):1170-1176.

[47] Diaz-Pinto A, Morales S, Naranjo V, Köhler T, Mossi JM, Navea A. CNNs for automatic glaucoma assessment using fundus images: An extensive validation. Biomedical Engineering Online. 2019;18(1):29.

[48] Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318(22):2211-2223.

[49] Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology. 2018;125(8):1199-1206.

[50] Li T, Gao Y, Wang K, et al. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences. 2019;501:511-

522.

[51] Abràmoff MD, Garvin MK, Sonka M. Retinal imaging and image analysis. IEEE Reviews in Biomedical Engineering. 2010;3:169-208.

[52] Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: The Messidor database. Image Analysis & Stereology. 2014;33(3):231-234.

[53] Yu J, Li X, Koh JY, et al. Vector-quantized image modeling with improved VQGAN. International Conference on Learning Representations. 2022.

[54] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:586-595.

[55] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017:1125-1134.

[56] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2009:248-255.

[57] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems. 2019;32:3347-3357.

[58] Ke A, Ellsworth W, Banez O, Narayanan AV, Callison-Burch C, Tseng E. Chextransfer: Performance and parameter efficiency of ImageNet models for chest x-ray interpretation. ACM Conference on Health, Inference, and Learning. 2021:116-124.

[59] Zhang B, Sennrich R. Root mean square layer normalization. Advances in Neural Information Processing Systems. 2019;32:12360-12371.

[60] Shazeer N. GLU variants improve transformer. arXiv preprint arXiv:2002.05202. 2020.

[61] Su J, Lu Y, Pan S, Murtadha A, Wen B, Liu Y. RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing. 2024;568:127063.

[62] Lu J, Clark C, Zellers R, Mottaghi R, Kembhavi A. Unified-IO: A unified model for vision, language, and multi-modal tasks. International Conference on Learning Representations. 2023.

[63] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361. 2020.

[64] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. Advances in Neural Information Processing Systems. 2022;35:30016-30030.

[65] Ho J, Salimans T. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598. 2022.

[66] Loshchilov I, Hutter F. Decoupled weight decay regularization. International Conference on Learning Representations. 2019.

[67] Micikevicius P, Narang S, Alben J, et al. Mixed precision training. International Conference on Learning Representations. 2018.

[68] Zhao Y, Gu A, Varma R, et al. PyTorch FSDP: Experiences on scaling fully sharded data parallel. Proceedings of the VLDB Endowment. 2023;16(12):3848-3860.

[69] Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. International Conference on Learning Representations. 2020.

[70] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems. 2017;30:6626-6637.

[71] Ansel J, Yang E, He H, et al. PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation. ACM ASPLOS Conference. 2024:929-943.

Note: Figure translations are in progress. See original paper for figures.