A Network Fitness Analysis of the Diffusion Model

Siting Chen(31520221154195), Lijiang Li(31520221154248), Zhiyu Yang(31520221154233), Jinlu Zhang(31520221154189)

Department of Artificial Intelligence, School of Information, Xiamen University

Abstract

Diffusion models can yield image sample quality superior to the current state-of-the-art generative models. However, most of the previous approaches use U-Net in the diffusion model and rarely employ a different network. We argue that it is necessary to try diverse network structures in the diffusion model, so that we can figure out the most suitable one. Based on this insight , We plan to perform experiments of image generation on three different neural networks: U-Net, Resnet and Transformer. Our study expects to propose a network fitness analysis of the diffusion model domain. We experiment with Cifar-10 in order to compare these methods and proved that the diffusion model based on U-Net works best.

Introduction

Image generation has always been a hot spot in computer vision and has a wide range of applications in real life. The target of image generation aims to synthesize realistic and vivid images through a generative model (Yan et al. 2016; Ma et al. 2018). GAN (Creswell et al. 2018; Karras, Laine, and Aila 2019; Karras et al. 2020) have dominated the field of image generation for a long period. By training two networks with adversarial loss, the generator is capable to synthesis striking image to treat the discriminator. However, GAN is prone to modal collapse (Thanh-Tung and Tran 2020; Srivastava et al. 2017; Liu et al. 2019), so other methods like VAE (Kingma and Welling 2013; Vahdat and Kautz 2020), autoregressive models (Oord et al. 2016; Van den Oord et al. 2016), flow-based models (Dinh, Sohl-Dickstein, and Bengio 2016; Kingma and Dhariwal 2018) and diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) have been developed to help release this dilemma in image generation, but their performance are still inferior to GAN models. Recently, a large text-to-image model named DALL·E (Ramesh et al. 2021) proposed by OpenAI is capable to synthesis striking and photo-realistic image conditioned to the input text, which arouse much more attention to the diffusion model since published. From then on, diffusion models have emerged as the new state-of-the-art in the field of image synthesis.

Specifically, a diffusion model can be regarded as a composition of forward noise-adding process and backward noise-reduction process (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Yang et al. 2022). In the forward diffusion process, by adding a very small amount of Gaussian noise to a real image in T steps, the data sample gradually loses its feature and is equivalent to an isotropic Gaussian distribution. And if the forward process could be reverse, a true sample can be recreated from a Gaussian noise input. Most of the diffusion model adopt a deep neural network to predict the added noise at each timestep, gradually denoising the input noise to a realistic image. This method is easy to implement (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020), since both forward and backward are deterministic Markov chains, the training of diffusion models is more stable than GANs. Not only does the diffusion model break the long-time dominance of GANs in image synthesis (Song and Ermon 2019; Song et al. 2020), they have shown potential in a variety of domains, ranging from computer vision (Amit et al. 2021; Baranchuk et al. 2021; Cai et al. 2020) to natural language processing (Austin et al. 2021a; Hoogeboom et al. 2021; Li et al. 2022b).

In order to generate better quality images, some sampling techniques have been applied in diffusion models (Ho and Salimans 2022; Dhariwal and Nichol 2021; Liu et al. 2021a; Kim, Kwon, and Ye 2022), while the network used in the noise prediction always keeps the same. Most methods used U-Net in the diffusion models and seldom adopted different networks. We believe it is necessary to try different network structures in diffusion model, as we can figure out the most suitable one by comparing the experiment results. We decide to conduct the experiments of image generation on three different neural networks: U-Net (Ronneberger, Fischer, and Brox 2015), Resnet (He et al. 2016) and Transformer (Vaswani et al. 2017; Devlin et al. 2018). These networks stand for three different paradigms in deep learning. U-Net has a typical encoder-decoder structure to perform feature extraction and reconstruction, which is widely used in medical image segmentation. Famous for its shortcut connection, Resnet successfully alleviate gradient vanishment or explosion. By adopting attention mechanism, transformer-based models advance state-of-the-art in many tasks in NLP, which also demonstrate promising results on certain vision tasks, specifically in classification and joint vision-language modeling. By conducting experiments on these three models, we look forward to proposing an analysis of the network adaptability in the field of diffusion model.

We conducted our experiments on Cifar-10 dataset and analyze the image generation results in both qualitative and quantitative perspectives. As for qualitative analysis, we compared the output of different models by visualizing the synthesized images. As for the quantitative analysis, we decide to use three metrics: FID, IS and accuracy to analyze the quality and diversity of the generated images. With our analysis and comparison, we could find that the diffusion model based on U-Net works best.

Related Work

Diffusion Model

As a class of deep generative models, diffusion model aims to transform the prior data distribution into random noise before revising the transformations step by step to rebuild a brand new sample with the same distribution as the prior. In the recent year, diffusion models have demonstrated remarkable results in the fields including computer vision (Nichol et al. 2022; Saharia et al. 2022; Zhang et al. 2022), nature language processing (Austin et al. 2021b; Gong et al. 2022), audio processing (Huang et al. 2022), interdisciplinary applications(Hoogeboom et al. 2022), etc.

Resnet

Resnet (He et al. 2016) is a deep neural networks with residual learning framework, proposed to learn residual of identity mapping, which can ease the training of networks. These technique are easier to optimize, and can gain accuracy from considerably increased depth. Resnet was first proposed to be applied in image recognition, but it widely used in many models as a backbone in other direction in computer vision (Li et al. 2022a; Wang et al. 2017).

U-Net

U-Net (Ronneberger, Fischer, and Brox 2015) relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture of U-Net consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. In image segmentation tasks, especially in medical image segmentation, U-Net is undoubtedly one of the most successful methods. There are many new convolutional neural network methods, but many still continue the framework of U-Net, like U-Net++ (Zhou et al. 2019), RelayNet (Roy et al. 2017), and so on.

Swin Transformer

Swin Transformer (Liu et al. 2021b) can serve as a general-purpose backbone for computer vsion. It is a hierarchical Transformer whose representation is computed with shifted windows, which scheme brings greater efficiency. This hierarchical architecture is flexible in modeling at various scales and has linear computational complexity with respect to image size.

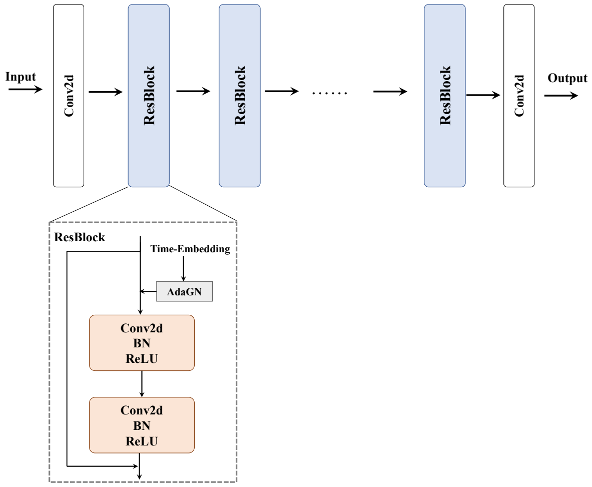


Figure 1: The architecture of Resnet used in DDPM.

Solution

Diffusion Model

Diffusion models are generative models that produce samples by inverting a corruption process. Given a data distribution $x_0 \sim q\left(x_0\right)$, diffusion models define a forward noising process q which produces latents x_1 through x_T by adding Gaussian noise at time t with variance $\beta_t \in (0,1)$ as follows:

$$q\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right) \tag{1}$$

therefore x_t can be obtained as follows:

$$x_t=\sqrt{1-\beta_t} x_{t-1}+\sqrt{\beta_t} \epsilon_t, \epsilon_t \sim \mathcal{N}(0, I) \tag{2}$$

Given a sufficiently large T and a well behaved schedule of β_t, the latent x_T is nearly an isotropic Gaussian distribution. Thus, if we know the mapping from x_t to x_{t-1}, we can we can sample $x_T \sim \mathcal{N}(0, I)$ and run the process in reverse to get a sample x_0. To achieve that, we used a neural network $\mu|\left(x_t\right)$ to approximate this mapping, which is trained to minimize the following loss function:

$$\left\|x_{t-1}-\mu\left(x_t, t\right)\right\|^2 \tag{3}$$

Alternatively, the network could also predict the noise ϵ. In this case, since $x_t=\sqrt{1-\beta_t} x_{t-1}+\sqrt{\beta_t} \epsilon_t$, we have:

$$x_{t-1}=\frac{1}{\sqrt{1-\beta_t}}\left(x_t-\sqrt{\beta_t} \epsilon_t\right) \tag{4}$$

Thus, $\mu\left(x_t, t\right)$ is rewritten as follows:

$$\mu\left(x_t, t\right)=\frac{1}{\sqrt{1-\beta_t}}\left(x_t-\sqrt{\beta_t} \epsilon_\theta\left(x_t, t\right)\right) \tag{5}$$

where $\epsilon_\theta\left(x_t, t\right)$ denotes the neural network and θ is its parameters. Then the loss function is:

$$\frac{\beta_t}{1-\beta_t}\left\|\epsilon_t-\epsilon_\theta\left(x_t, t\right)\right\|^2 \tag{6}$$

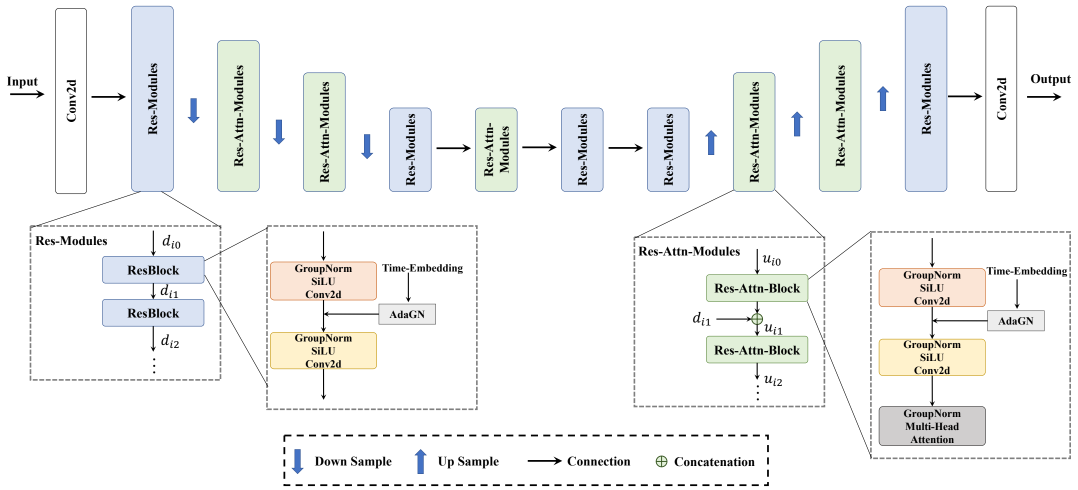


Figure 2: The architecture of U-Net used in DDPM.

After the network is well trained, we can obtain x_{t-1} from x_t as follows:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(x_t - \sqrt{\beta_t} \epsilon_\theta \left(x_t, t \right) \right) + \sigma_t z \tag{7}$$

where $z \sim N(0, I)$.

A typical diffusion model takes a image sample x_{t-1} and a timestep t as input and predicts noise ϵ_{t-1} as output, where x_{t-1} can be obtained by Equation (2). Therefor, the process of inverse diffusion model can be estimated as Equation (8),

$$G_\Theta(x_{t-1}, t) = \epsilon_{t-1} \tag{8}$$

where G can be a deep network and Θ is the model parameters that need to optimized. As described in the work of (Ho, Jain, and Abbeel 2020), t will be encoded by a MLP and added in the process of transition. We will describe in details about how we apply three different nerual networks in DDPM and their corresponding structure.

Resnet

The adaptation of Resnet is simple. In Resnet, a ResBlock can be designed as Equation (9).

$$\mathbf{y} = \mathcal{F}\left(\mathbf{x}, \{W_i\}\right) + \mathbf{x} \tag{9}$$

To fuse the time information into the Resnet-based diffusion model, we makes some modifications. As shown in Figure 1, we simply add the projection of timestep embedding at the begginning of each ResBlock, after which two identical block of convolution layer with Batch Normalization and ReLU activation is performed on the input feature. This can be formulated as Equation (10).

$$\mathbf{y} = \mathcal{F}\left(\mathbf{AdaGN(t)}, \mathbf{x}, \{W_i\}\right) + \mathbf{x} \tag{10}$$

In addition, we apply AdaGN on the input timestep embedding, which is described as Equation (11).

$$AdaGN(h, y) = y_s GroupNorm(h) + y_b \tag{11}$$

U-Net

U-Net(Ronneberger, Fischer, and Brox 2015) model is consist of a stack of upsampling layers and downsampling layers, with skip connections connecting the layers with the same spatial size. (Dhariwal and Nichol 2021) confirmed that making some changes to the U-Net architecture can further improved performance on the CIFAR-10 and CelebA-64 datasets. So the final architecture of U-Net we apply in this paper is illustrated as Figure 2. The U-Net model is consist of a stack of residual layers and downsampling convolutions, followed by a stack of residual layers with upsampling convolutions, and the skip connections is the same of (Ronneberger, Fischer, and Brox 2015). We use a global attention layer at the 16 × 16 resolution with a single head. In the dowsampling process, a projection of the timestep embedding will be added between two groups, each of which contains a convolution layers followed by Group Normailization and SiLU activation in a ResBlock. In the upsampling process, timestep embedding is feeded after AdaGN in each Res-Attn-Block.

Swin Transformer

The architecture of Swin Transformer (Liu et al. 2021b) we use in DDPM is illustrated as Figure 3. We follow the work of the original Swin Transformer but make somd adjustion. In (Liu et al. 2021b), each Swin Transformer block consists of a shifted window based MSA module, followed by a 2-layer MLP with GELU nonlinearity in between. A LayerNorm (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module. To facilitate Swin Transformer to the diffusion model, we add the timestep embedding in each Swin Transformer layer after the MSA module and apply AdaGN illustrated as Equation (11).

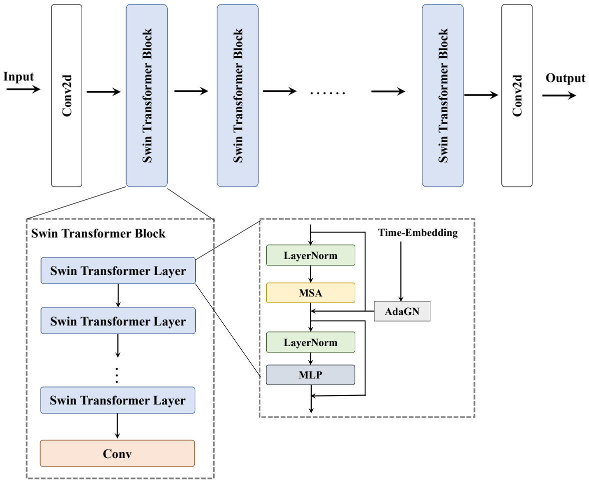


Figure 3: The architecture of Swin Transformer used in DDPM.

Optimization

Equation (12) is the original objective function, which is derived by considering the variational lower bound.

$$\begin{aligned} \mathbb{E}\left[-\log p_\theta\left(\mathbf{x}_0\right)\right] &\leq \mathbb{E}_q\left[-\log \frac{p_\theta\left(\mathbf{x}_{0:T}\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right] \\ &= \mathbb{E}_q\left[-\log p\left(\mathbf{x}_T\right)-\sum_{t \geq 1} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right] \end{aligned} \tag{12}$$

In order to simplify the training process, we follow (Ho, Jain, and Abbeel 2020) and use Equation (13) to train the three networks.

$$L_{\text{simple }}=E_{t, x_0, \epsilon}\left[\left\|\epsilon-\epsilon_\theta\left(x_t, t\right)\right\|^2\right] \tag{13}$$

Experiment

Extensive experimentation is carried out to evaluate the proposed solutions. We study three different network structures mentioned above of diffusion model respectively.

Experiment Setup

Our method is implemented in PyTorch 1.7.1. All the models in the experiments were trained on 1 NVIDIA GeForce RTX 3090 GPU. All the models were trained with Adamw optimizer with betas = (0.9, 0.999), lr = 0.0001 and weight-decay = 0. The number of iterations of training is 50w for all models. Follows Improved DDPM (Nichol and Dhariwal 2021), the diffusion steps (T) is set as 4000. In addition, we used cosine noise schedule (Nichol and Dhariwal 2021), which means that the β_t will change with time t as follows:

$$\begin{aligned} \beta_t &= 1-\frac{\overline{\alpha}_t}{\overline{\alpha}_{t-1}} \\ \overline{\alpha}_t &= \frac{f(t)}{f(0)} \\ f(t) &= \cos\left(\frac{t/T+s}{1+s}\cdot\frac{\pi}{2}\right) \end{aligned} \tag{14}$$

where s is a small offset to prevent β_t from being too small near $t=0$.

Datasets

We compare image generation performance on CIFAR-10 (Krizhevsky, Hinton et al. 2009), a subset of the Tiny Images dataset and consists of 60000 32×32 color images. The images consist of 10 mutually exclusive classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. There are 6000 images per class with 5000 training and 1000 testing images per class. All compared models are trained on this dataset.

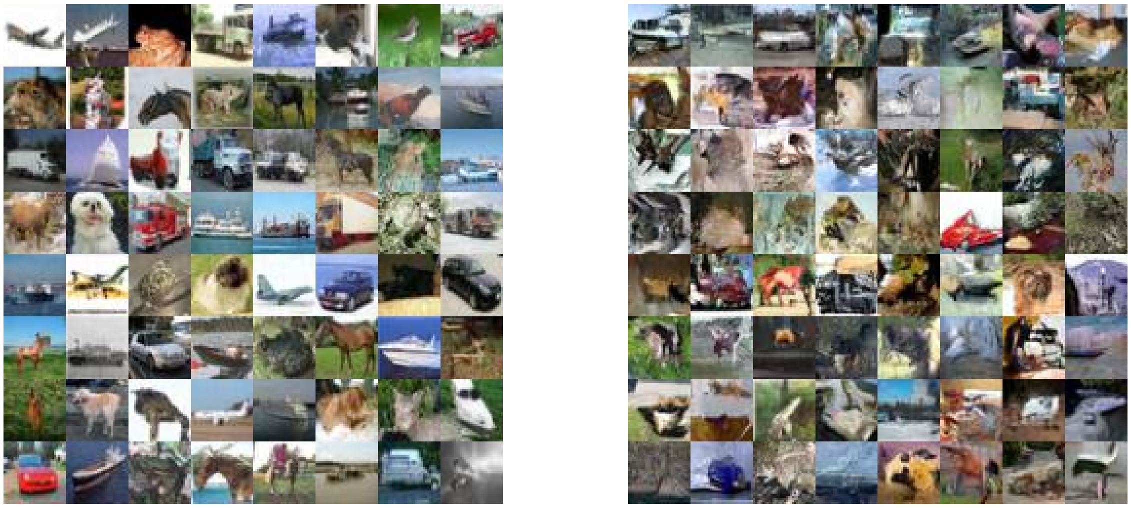
Evalution Metrics

The visual quality of generated or manipulated images is evaluated through the widely-used FID (Heusel et al. 2017) metrics and Inception Score (IS) (Li et al. 2019). These two are regarded as typical metrics for evaluating the model for image generation. FID measures the distance between two sets of images, computed by the mean value and covariance of the generated images and the original images. The smaller the difference, the better the generated image is. IS pays particular attention to the diversity and clarity of the resulting images. Higher IS results in better image quality.

$$\begin{aligned} \operatorname{FID}(Y, \hat{Y}) &= \left\|\mu_Y-\mu_{\hat{Y}}\right\|_2^2 \\ &+\operatorname{tr}\left(\Sigma_Y+\Sigma_{\hat{Y}}-2\left(\Sigma_Y \Sigma_{\hat{Y}}\right)^{\frac{1}{2}}\right) \end{aligned} \tag{15}$$

where $(\mu_Y, \sum_Y)$ is generate image set, $\left(\hat{\mu_Y}, \hat{\Sigma_Y}\right)$is original image set.

We first quantitatively evaluate three network structures, the U-Net, SwinIR and Restnet. The results are shown in Table 1. We measure the results per 50000 training epochs. It is evident that the model based on U-Net outperforms the other models, achieving FID of 9.81 at 25 w epochs. The result of SwinIR is lower than that of U-Net, and the result reaches fid of 40.91 at 40w epochs. This suggests that SwinIR does not have a downsampling and upsampling structure, which may lead to lack of global information of the whole images. Resnet achieves FID on CIFAR-10 over 455, shows obviously that it is difficult to converge, we suggest possibly results of the lack of attention mechanism. The U-Net based model achieves inception score of 9.759, significantly outperforming the SwinIR based network and Resnet based model with inception score of 5.978 and 1.209, respectively. The U-Net based model achieves both high accuracy and Recall. The U-Net Accuracy is increased from 0.5498 to 0.6316



(a) Generated by U-Net based.

(b) Generated by SwinIR based.

Figure 4: The images generated by U-Net based and SwinIR model.

	epochs	FID (numsample=50000) ↓	IS (numsample = 50000) ↑	Precision ↑	Recall ↑
U-Net	25w	9.81	9.809	0.6286	0.5418
	40w	11.91	9.843	0.6269	0.5154
	50w	13.52	9.759	0.6316	0.4961
SwinIR	25w	48.79	5.314	0.5498	0.3465
	40w	43.56	5.734	0.5394	0.3806
	50w	40.91	5.978	0.5347	0.4033
ResNet	25w	455.28	1.209	0	0
	40w	455.42	1.210	0	0
	50w	455.35	1.209	0	0

Table 1: Experiment Result

when compared to SwinIR, as well as Recall from 0.4033 to 0.5418. U-Net performs better than the other two ones, then we will give further analysis of the generated images. In the Figure 4,we qualitatively examine the images generated by U-Net based and SwinIR models. We don't show images generated by Resnet based models because it can't converge. Comparing with SwinIR, The images generated by U-Net based diffusion model are clear and more realistic with smooth lines, but the result is still not perfect, which leaves room to improve.

Conclusion

In this paper, we perform experiments of image generation on three different neural networks: U-Net, Resnet and Transformer. The results show that U-Net outperforms the others, achieving FID on CIFAR-10 at 9.81. The result of SwinIR is lower than that of U-Net, possible reason is that SwinIR does not have a downsampling and upsampling structure. Resnet is difficult to converge. The U-Net based model can generate better quality images. Our experiments indicate that compares to Resnet and Transformer, U-Net is the most suitable one for diffusion model.

References

Amit, T.; Nachmani, E.; Shaharbany, T.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and van den Berg, R. 2021a. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34: 17981–17993.

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and van den Berg, R. 2021b. Structured Denoising Diffusion Models in Discrete State-Spaces. *arXiv: Learning*.

Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrulkov, V.; and Babenko, A. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*.

Cai, R.; Yang, G.; Averbuch-Elor, H.; Hao, Z.; Belongie, S.; Snavely, N.; and Hariharan, B. 2020. Learning gradient fields for shape generation. In *European Conference on Computer Vision*, 364–381. Springer.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2022. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hoogeboom, E.; Nielsen, D.; Jaini, P.; Forré, P.; and Welling, M. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34: 12454–12465.

Hoogeboom, E.; Satorras, V. G.; Vignac, C.; and Welling, M. 2022. Equivariant Diffusion for Molecule Generation in 3D. *international conference on machine learning*.

Huang, R.; Zhao, Z.; Liu, H.; Liu, J.; Cui, C.; Ren, Y.; and Prodiff. 2022. ProDiff: Progressive Fast Diffusion Model for High-Quality Text-to-Speech.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2426–2435.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, B.; Wang, C.; Reddy, P.; Kim, S.; and Scherer, S. 2022a. Airdet: Few-shot detection without fine-tuning for autonomous exploration. In *European Conference on Computer Vision*, 427–444. Springer.

Li, W.; Zhang, P.; Zhang, L.; Huang, Q.; He, X.; Lyu, S.; and Gao, J. 2019. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12174–12182.

Li, X. L.; Thickstun, J.; Gulrajani, I.; Liang, P.; and Hashimoto, T. B. 2022b. Diffusion-LM Improves Controllable Text Generation. *arXiv preprint arXiv:2205.14217*.

Liu, K.; Tang, W.; Zhou, F.; and Qiu, G. 2019. Spectral regularization for combating mode collapse in GANs. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6382–6390.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2021a. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Ma, L.; Sun, Q.; Georgoulis, S.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 99–108.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *international conference on machine learning*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Roy, A. G.; Conjeti, S.; Karri, S. P. K.; Sheet, D.; Katouzian, A.; Wachinger, C.; and Navab, N. 2017. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical optics express*, 8(8): 3627–3642.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Kamyar, S.; Ghasemipour, S.; Karagol, B.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Srivastava, A.; Valkov, L.; Russell, C.; Gutmann, M. U.; and Sutton, C. 2017. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30.

Thanh-Tung, H.; and Tran, T. 2020. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, 1–10. IEEE.

Vahdat, A.; and Kautz, J. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33: 19667–19679.

Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; et al. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Yan, X.; Yang, J.; Sohn, K.; and Lee, H. 2016. Attribute2image: Conditional image generation from visual attributes. In *European conference on computer vision*, 776–791. Springer.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Shao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2022. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6): 1856–1867.