Multi-Emotional Expression Generation from a Single Facial Expression

Yansong Liu 30920241154569, Mingyi Xu 30920241154580, Yuhang Lin 30920241154568

School of Informatics, Xiamen University
{30920241154569,30920241154580,30920241154568}@stu.xmu.edu.cn

Abstract

Expressions are crucial carriers of human emotions, with their subtle variations often conveying complex feelings. Accurately capturing the relationship between expressions and emotions, and generating images that can convey delicate emotions, poses significant challenges in the field of image generation. This paper proposes a framework for expression generation based on diffusion models, aiming to enhance the quality of generated images and the accuracy of emotional expression. To validate the effectiveness of this framework, we conducted comparative experiments using three mainstream generative models: StarGAN, VAE, and Diffusion. Using the same input data, we evaluated the quality of the generated results by calculating the Fréchet Inception Distance (FID). We also conducted an in-depth analysis of the advantages and disadvantages of each model, exploring their differences in generative diversity, training stability, and expression authenticity.

Introduction

Facial expressions serve as a crucial medium for human emotions, with their subtle variations conveying rich emotional information. In the field of image generation, the ability to accurately express a range of emotions from a single expression is significant for enhancing the naturalness of human-computer interaction and deepening our understanding of human emotions. With the advancement of deep learning technologies, utilizing methods such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models for generating expression images has become a hot research topic. Notably, diffusion models have recently demonstrated enormous potential in image generation tasks, with their advantages in preserving image details and quality gradually being recognized.

Previous studies have explored the use of VAEs and GANs to generate facial expression images with various emotional representations. For example, latent space interpolation through VAEs allows for smooth transitions between expressions, and emotion-specific GANs enable detailed control over the emotional features of generated expressions. However, these methods still face challenges in improving the diversity of generated expressions and the authenticity of the conveyed emotions. Moreover, the complexity of expression generation tasks poses a challenge in effectively capturing and expressing the multiple emotions underlying a single expression.

This study aims to explore a framework for expression generation based on diffusion models to enhance the accuracy and diversity of emotional expression in generated images. We selected three mainstream generative models—StarGAN, VAE, and diffusion models—for comparative research, intending to analyze the performance differences of each model in generating expression images through comparative experiments. We believe that an in-depth investigation of the strengths and weaknesses of these models will provide valuable insights for future research on expression generation.

2 Relatedwork

2.1 Image Generation

Image Generation is a type of task that converts one type of image into another, typically requiring the preservation of the original image's content while generating a new image with certain specific features or styles. At its core, it learns the mapping relationship between pairs of images (i.e., ”input image–target image”) to enable transformations from one image domain to another. This task has wide applications in various image processing fields, such as style transfer, semantic segmentation, expression generation, and image enhancement. The primary goal of this project is to generate diverse expressions from a single facial expression image. The study of generating multiple emotional expressions from a single facial expression has garnered significant attention in the fields of computer vision and affective computing. Popular generative models for image generation include GAN, VAE, and Diffusion.

2.2 StarGAN

StarGAN is an extension of Generative Adversarial Networks (GANs) designed to achieve multi-domain image generation with a single model. Unlike traditional GAN models, StarGAN utilizes a structure comprising a generator and a discriminator but can handle images from multiple input domains, enabling the transformation between different emotions and styles. By incorporating conditional information, StarGAN learns the relationships between different emotional states without requiring a large amount of labeled data, thus facilitating high-quality image generation. The innovation of StarGAN lies in its ability to use the same generator network for transformations across various emotional states, eliminating the need to train a separate model for each emotion. This feature significantly enhances the efficiency and flexibility of the generative model. Numerous studies have explored the applications of StarGAN in emotional synthesis. For example, Choi et al. (2018) first proposed the StarGAN framework, which was utilized for multi-domain image translation, demonstrating the capability of generating different emotional expressions by manipulating input conditions. Zhang et al. (2019) further advanced StarGAN by introducing an improved training method, resulting in a model that performs more stably and accurately when generating expressions with nuanced emotional characteristics.

2.3 VAE

Variational Autoencoders (VAEs) are probabilistic generative models composed of an encoder and a decoder. The encoder maps input images to a continuous latent space, while the decoder samples from this latent space to generate images. The advantage of VAEs is their ability to produce images with diversity and continuity, making them suitable for generating images with controlled features. The framework proposed by Kingma and Welling (2014) offers a probability-based solution to generative modeling, which has been widely applied to expression generation and emotion recognition tasks. In the field of emotion generation, the latent space of VAEs can capture rich facial expression features, allowing for smooth interpolation from a single expression to various emotions. Recent work by Tewari et al. (2020) further demonstrates VAEs' effectiveness in generating diverse emotional expressions by modeling facial features within the latent space, enabling the exploration of multiple emotional states based on a base expression. This approach not only allows for natural transitions between expressions but also preserves the structural characteristics of the original expression. Additionally, VAEs effectively capture and produce complex emotional features, allowing for nuanced emotional expressions that align with intuitive human perception of emotions.

2.4 Diffusion

Diffusion Models have gained significant attention as powerful generative tools due to their ability to generate high-quality, stable images. The generative process in diffusion models involves two stages: adding noise gradually to an image and then denoising it to reconstruct a clear image. Sohl-Dickstein et al. (2015) initially introduced the principles of diffusion models by simulating a gradual denoising path for image generation. In subsequent research, improvements by Ho et al. (2020) showcased the model's potential for generating complex, diverse outputs, making it especially suited for complex image generation tasks. In the task of emotion generation, diffusion models present unique advantages, allowing for subtle emotional variations based on a single initial expression. This gradual transformation process accurately captures emotional features and maintains structural integrity in the generated images throughout the emotion generation process. Therefore, in multi-emotion generation tasks, diffusion models provide an effective solution, resulting in smoother emotional transitions. Compared to other generative methods, the denoising process in diffusion models significantly reduces noise and artifacts in generated images, improving both the quality and stability of the final outputs.

3 Method

3.1 Conditional Convolutional VAE

The Conditional Convolutional VAE is an extension of the traditional Variational Autoencoder, incorporating convolutional layers and conditioning mechanisms to handle image data and emotion labels effectively. This architecture is particularly suited for tasks such as facial expression synthesis, where both spatial information and conditional attributes are crucial.

Model Architecture: Our Conditional VAE consists of an encoder and a decoder, both conditioned on emotion labels to guide the synthesis process.

Encoder: The encoder comprises a series of convolutional layers that extract hierarchical features from the input image. These features are flattened and concatenated with a one-hot encoded emotion label vector. This combined representation is used to compute the parameters of the latent distribution, specifically the mean μ and the log variance $\log \sigma^2$.

Decoder: The decoder reconstructs the image from the latent vector, which is concatenated with the emotion label. It uses transposed convolutional layers to upsample the latent representation back to the original image size, ensuring that the generated facial expression aligns with the specified emotion.

$$h = ReLU(Conv2d(x)) \tag{1}$$

$$z_{input} = Flatten(h) \oplus c \tag{2}$$

$$\mu, \log \sigma^2 = Linear(z_{input}) \tag{3}$$

The reparameterization trick is employed to allow gradient descent through the stochastic latent space, with the latent variable z computed as:

$$z = \mu + \epsilon \cdot \sigma, \quad \epsilon \sim \mathcal{N}(0, I)$$

Loss Function: The Conditional VAE is trained to minimize a combined loss function comprising reconstruction and KL divergence losses:

Reconstruction Loss: Ensures the generated image closely resembles the input image in terms of pixel-wise similarity, measured by Mean Squared Error (MSE).

$$\mathcal{L}_{recon} = MSE(x, \hat{x})$$

KL Divergence Loss: Regularizes the latent space to follow a standard normal distribution, promoting smooth and continuous latent representations.

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{i=1}^{D}(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2)$$

The total loss is given by:

$$\mathcal{L} = \mathcal{L}_{recon} + \beta \cdot \mathcal{L}_{KL}$$

where β is a hyperparameter that balances the two components.

Training Efficiency: The use of a single encoder-decoder pair conditioned on emotion labels allows efficient training across different expression domains. This approach reduces model complexity compared to training separate models for each expression type, enhancing both flexibility and scalability.

This architecture enables our Conditional VAE to generate diverse facial expressions while maintaining high fidelity to the input image, making it a powerful tool for facial expression synthesis tasks.

3.2 StarGAN

StarGAN is a multi-domain image generation framework based on Generative Adversarial Networks (GANs), designed to efficiently perform image translation across different domains. It demonstrates strong versatility and can be widely applied to various image generation tasks, such as facial expression transfer, style transfer, and attribute editing. StarGAN is an extension of CycleGAN, inheriting the cycle-consistency property and utilizing adversarial loss to learn mappings between different domains (e.g., from x to y and vice versa). The key difference from CycleGAN lies in the use of a single generator to perform bidirectional mapping, instead of two generators in a cyclic structure. Specifically, StarGAN introduces a reconstruction loss to ensure feature consistency between the input image x and the generated image G(x), which is defined as:

$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}\left[\|x - G(G(x,c),c')\|_1\right] \tag{4}$$

Here, the generator G takes the generated image G(x,c) and the domain label c' of the original image as inputs and attempts to reconstruct the original image x. This approach reduces the number of model parameters, improving the training efficiency, while enabling the generator to learn the bidirectional mappings x to y and y to x.

A distinctive feature of StarGAN is the introduction of the mask vector m, which facilitates joint training on multi-domain datasets. During training with multiple domains, StarGAN uses a unified label vector c to represent the target domains, which is represented as:

$$\tilde{c} = [c_1, \ldots, c_n, m], \tag{5}$$

where each domain label c_i is represented as a one-hot vector. The mask vector m is an n-dimensional one-hot vector that indicates the target domain to which the generator and discriminator should translate the input image.

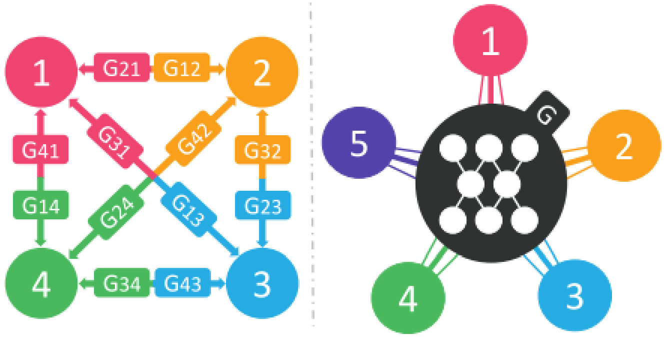

Figure 1: Comparison Between Traditional Cross-Domain Models(left) and StarGAN(right)

This design sets StarGAN apart from traditional cross-domain models, as it does not require training separate generators for each pair of domains. Instead, a single generator is used to learn mappings across multiple domains. This significantly simplifies the model architecture and improves training efficiency, making StarGAN both more efficient and flexible for multi-domain image generation tasks.

3.3 Diffusion

3.3.1 Diffusion Model

Diffusion models are a class of probabilistic generative models that focus on modeling the gradual denoising process of data to generate target samples. Specifically, a diffusion model consists of a forward process and a reverse process. In the forward process, Gaussian noise is progressively added to the original data until it approaches a pure noise distribution. In the reverse process, the model learns to iteratively denoise the data, gradually restoring it from pure noise to the original data distribution. This step-by-step approach enables diffusion models to generate high-quality and realistic images. Compared to traditional Generative Adversarial Networks (GANs), diffusion models offer greater training stability and enhanced diversity in generated samples, though the generation process is often more time-consuming. In recent years, diffusion models have achieved remarkable results in tasks such as image and video generation, making them a prominent focus in the field of generative modeling.

Inspired by non-equilibrium thermodynamics, the diffusion model now produces the most advanced image quality,with examples as follows:

The training process of diffusion models involves two main phases: the forward process and the reverse process. In the forward process, noise is progressively added to the input data until it is completely transformed into noise. This process is typically controlled by a predefined noise scheduler, which gradually increases the amount of noise at each step. In the reverse process, the model's objective is to learn how to recover the original data from the noise. To achieve this, the model uses a neural network to predict the amount of noise to remove at each denoising step, gradually restoring the data to its original distribution. During training, a

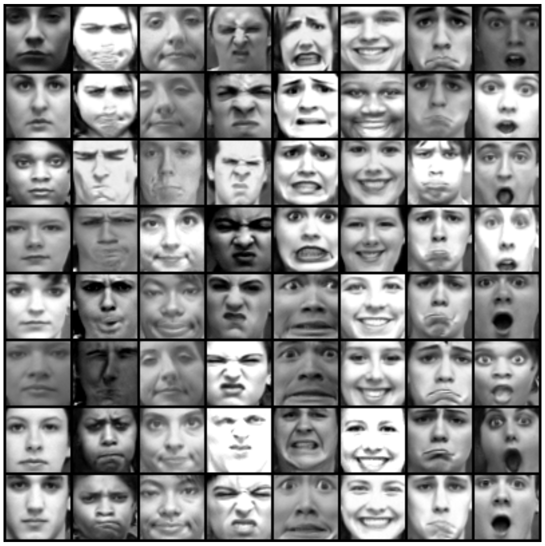

Figure 2: Diffusion Model Generation Examples

loss function is used to measure the prediction error at each denoising step, with the most common loss function being Mean Squared Error (MSE). The goal of training is to minimize these errors, enabling the model to efficiently recover the data from noise and generate high-quality images.The entire process is shown in Figure 3.

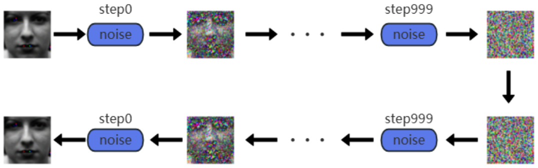

Figure 3: Diffusion Model Working Principle

3.3.2 Conditional Diffusion Model

To further enhance the performance of diffusion models, researchers have introduced Conditional Diffusion Models. Conditional diffusion models incorporate conditional information (such as labels, image features, or other types of auxiliary data) to guide the generation process. This enables the model not only to generate realistic images but also to generate images that meet specific conditions. For example, a conditional diffusion model can generate images with specific emotional expressions based on given emotion labels or generate images that match a textual description. By introducing such control, conditional diffusion models offer greater flexibility and customization in the generation process.

To tackle the task of facial expression generation, we aim to leverage the powerful reconstruction and manipulation capabilities of diffusion models. A standard diffusion model, as illustrated in Figure 3, takes an image as input x_0 and iteratively adds Gaussian noise at configurable timesteps T, generating a series of noisy inputs $x_1, \ldots, x_T$. Ideally, x_T follows a Gaussian distribution, appearing as random noise. Starting from the noisy image, a denoising network predicts the added noise at a given timestep, i.e., $p(x_{t-1}|x_t)$. Typically, the denoising network is trained using a mean squared error (MSE) loss between the actual added noise and the predicted noise.With this setup, during inference, a user can sample random noise at timestep T, iteratively use the denoising network to predict the added noise, and remove it step-by-step until reaching the first timestep. At this point, the output is an image generated by the diffusion model.The conditional diffusion model has the same overall architecture, except that the denoising network takes an additional ”condition” input along with the noisy input x.

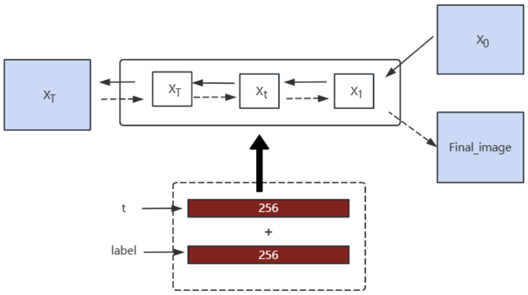

Figure 4: Model Architecture

In our model, the denoising network condition includes an embedding of the timestep information and the output label semantics, with emotional labels projected into a 256-dimensional space using an embedding layer. The additional label embedding is designed to guide the denoising process by emphasizing the desired target facial expression in the output image.

As mentioned earlier, the training process is supervised by the MSE loss between the predicted noise and the actual noise. Specifically, during training, the labels passed to the semantic embedding layer correspond to the facial expression labels of the input image. During inference, we pass the label of the desired emotion, and the model generates the target facial expression in the output image.

3.3.3 Loss function of Model

During the training of the Conditional Diffusion Model, there are forward noise addition and backward denoising processes. The denoising network is trained using the mean squared error (MSE) loss between the actual added noise and the predicted noise.

$$\mathcal{L}_{denoise} = MSE(x, \hat{x})$$

4 Experiments

4.1 Implementation Details

4.1.1 Training Conditional VAE

The Conditional VAE model was trained using the CK+ dataset, which includes a collection of facial expression images with corresponding emotion labels. The dataset was preprocessed to focus on facial regions, ensuring each image was centered and normalized. Images were augmented to maintain consistency in the training data, with each image resized to 64×64 pixels, aligning with the model's input requirements.

The model was implemented using the PyTorch framework. Training utilized the Adam optimizer with a learning rate of 0.001. The optimizer's parameters were set to default values, providing stability and efficient convergence during training. The model was trained on a single NVIDIA GPU, employing a batch size of 64 to balance memory usage and computational efficiency. The training spanned 100,000 epochs, with model checkpoints saved every 500 epochs to facilitate progress monitoring and potential recovery.

The loss function incorporated both mean squared error (MSE) and a Kullback-Leibler (KL) divergence term, with a weighting factor of $\beta = 0.001$ applied to the KL divergence. This configuration aimed to ensure a balance between reconstruction fidelity and latent space regularization. The training process was monitored using TensorBoard, which provided real-time insights into loss dynamics and model performance, aiding in the fine-tuning of model parameters and training strategies.

4.1.2 Training StarGAN

StarGAN was trained on the CK+ dataset, which contains 10,708 facial expression images categorized into seven emotion labels: angry, contempt, disgust, fear, happy, neutral, sadness, and surprise. Due to variations in facial positioning within the dataset, preprocessing was performed prior to training. This involved centering the faces and resizing the images to 224×224 pixels to ensure consistency in input data.The model was implemented using the PyTorch framework. During training, the Adam optimizer was employed with a learning rate of 0.0001 and momentum parameters set to β_1=0.5 and β_2=0.999. The training was conducted on a single RTX 3090 GPU with a batch size of 16, taking approximately 5 hours to complete.This training configuration and optimization setup provided a solid foundation for StarGAN's performance in facial expression generation tasks.

4.1.3 Training Conditional Diffusion Model

The Conditional Diffusion Model was trained using the CK+ dataset, which includes facial expression images with corresponding emotion labels. The dataset was preprocessed, and each image was resized to 48×48 pixels.

The model was implemented using the PyTorch framework. During training, the AdamW optimizer was used with a learning rate of 0.0003 and a weight decay coefficient of 0.001 to ensure stability and efficient convergence. The model was trained on a single NVIDIA 3090 GPU, with a batch size of 32 to balance memory usage and computational efficiency. The training spanned 1000 epochs, with model checkpoints saved every 10 epochs to monitor progress and allow for potential recovery.

The loss function utilized mean squared error (MSE), and the denoising network was trained by computing the MSE loss between the actual added noise and the predicted noise.

4.1.4 Qualitative Comparison

We trained the model using the same dataset and evaluated it on the same test set to generate images corresponding to seven facial expressions. The results were visually compared to assess the quality of the generated outputs.

In Figure 5, we can observe the facial expressions generated by VAE, StarGAN, and Diffusion models. Overall, the images generated by VAE exhibit relatively low facial clarity in expression generation. StarGAN generates images that maintain a high degree of similarity to the original images, particularly in the upper facial regions. However, in the lower facial regions, especially around the mouth, a phenomenon known as ”texture sticking” occurs, leading to noticeable spatial distortions of features like teeth or lips in the generated images. The Diffusion model designed in this study struggles to ensure that the generated faces resemble the input faces, but it does exhibit the best expression representation among the three models.

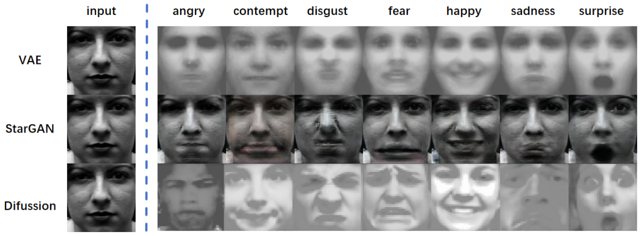

Figure 5: Comparison of Expression Generation Samples from Different Models

Furthermore, the lip features for generating other expressions are more complex and variable, making them difficult for the models to fully capture, and the results for all three models are not ideal. The limited size of the dataset may also contribute to this issue.

5 Conclusion

This paper addresses the problem of generating multiple facial expressions from a single image. We approach this issue using state-of-the-art image generation algorithms, including VAE, StarGAN, and Diffusion models. Using the same dataset as input, we evaluate the performance of these methods through both quantitative and qualitative metrics. Ultimately, we find that StarGAN performs the best for this task. This contrasts with the best performance typically achieved by diffusion models, leading us to hypothesize that modifying a basic diffusion framework similar to StarGAN may yield better results. This is something we plan to explore further in the future.

References

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096*.

Cao, H.; Tan, C.; Gao, Z.; et al. 2022. A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646*.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020a. Generative Pretraining from Pixels. In *International Conference on Machine Learning*.

Chen, N.; Zhang, Y.; Zen, H.; Weiss, R. J.; Norouzi, M.; and Chan, W. 2020b. WaveGrad: Estimating Gradients for Waveform Generation. *arXiv preprint arXiv:2009.00713*.

Child, R. 2021. Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. *arXiv preprint arXiv:2011.10650*.

Dayan, P.; Hinton, G. E.; Neal, R. M.; and Zemel, R. S. 1995. The Helmholtz Machine. *Neural Computation*, 7(5): 889–904.

Du, Y.; and Mordatch, I. 2019. Implicit Generation and Generalization in Energy-based Models. *arXiv preprint arXiv:1903.08689*.

Dumoulin, V.; Shlens, J.; and Kudlur, M. 2017. A Learned Representation for Artistic Style. *arXiv preprint arXiv:1610.07629*.

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

Huang, X.; Li, Y.; Poursaeed, O.; Hopcroft, J.; and Belongie, S. 2017. Stacked Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*.

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*, 8107–8116.

Lu, C.; Yu, C.; Song, Y.; and Ermon, S. 2022. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, 234–241. Springer International Publishing.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.

Wang, Z.; Zheng, H.; He, P.; et al. 2022. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*.

Zhao, Z.; Singh, S.; Lee, H.; et al. 2021. Improved consistency regularization for GANs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11033–11041.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*.