Enhancing Text-to-Image Synthesis with CLIP-Integrated Diffusion Models

Huayu Zeng 36920241153276[1], Jiahao Lu 36920241153242[1], Jiahao Gu 23020241154393[2]
Fangkun Chen 23020241154369[2], Boyang Wu 23020241154454[2]

[1]Institute of Artificial Intelligence, Xiamen University [2] School of Informatics, Xiamen University

Abstract

In recent years, the field of Artificial Intelligence Generated Content (AIGC) has made significant strides, particularly in image generation and editing. Diffusion Models generate high-quality text-to-image synthesis by simulating a progressive diffusion process that introduces noise into the data, followed by a denoising process that recovers the data from the noise, ultimately generating images with rich details and diversity based on textual descriptions. However, Diffusion Models face challenges such as the difficulty of aligning semantics between text and images, inconsistencies and lack of diversity in generated results, and limited contextual understanding. To address these issues, we introduce **CLIP (Contrastive Language-Image Pre-training)** to collaborate with Diffusion Models. This integration aims to enhance the quality and diversity of image generation, improve the model's generalization ability, and achieve more efficient and stable image editing outcomes.

Introduction

AIGC text-to-image technology is currently undergoing a surge in growth and garnering extensive attention. Diffusion Models, a class of deep generative models predicated on probabilistic processes, simulate the diffusion process by incrementally introducing noise into data and then employ a reverse diffusion process to systematically remove this noise, thereby restoring the original data. In image generation, these models begin by transforming a crisp image into pure noise and, through the learned reversal process, are capable of reconstructing the image from the noise. This approach is adept at producing high-quality images and excels in handling complex data distributions, which has propelled Diffusion Models to the forefront of text-to-image synthesis applications. They can generate images that are rich in detail and diverse, aligning closely with textual descriptions. With the impetus of Diffusion Models, text-to-image technology has reached a new plateau in delivering high-quality image generation services. The evolution of this domain has not only catalyzed a novel creator economy but also foreshadows the trajectory of future content creation.

To date, Diffusion Models have made significant strides in the field of image generation, surpassing the previous state-of-the-art, which was held by Generative Adversarial Networks (GANs). As technology continues to evolve, Diffusion Models have not only matched but exceeded the capabilities of GANs in the task of image generation, marking a new era in the realm of artificial intelligence and machine learning.

Figure 1: Text-driven facial modifications

While Diffusion Models have achieved notable advancements in image generation, they are not without their challenges and constraints. In this paper, we mainly focus on three challenges:

(1) Semantic Alignment between Text and Image. In diffusion models, the process of generating images primarily involves iteratively removing noise from random noise to construct the image. This process is fundamentally a "pixel-level" reconstruction of the image. However, in the absence of CLIP, diffusion models lack a direct mechanism to ensure strong semantic alignment between the generated image and the input textual description.

(2) The inconsistency in generated results and lack of diversity. Diffusion models generate images by progressively denoising, a process that heavily relies on the initial noise and the learned distribution. In the absence of CLIP to establish a more precise connection between the text and the

image, the diffusion model may fail to accurately capture the implicit details embedded in the text.

(3) Limited contextual understanding of the model. Diffusion models typically struggle to deeply comprehend the intricate contextual and situational nuances embedded within textual descriptions. As a result, they often face challenges in generating images that align well with the specific context or situational requirements outlined in the text. This limitation leads to the generation of results that lack both naturalness and accuracy, as the models are unable to capture the complex interdependencies between various elements within the given context. Consequently, the generated images may fail to fully reflect the intended meaning or specific details described in the text, compromising the overall quality and coherence of the output.

In brief, Diffusion Models face challenges in perceiving context and environment during the text-to-image generation process. Their limited ability to fully grasp the intricacies of textual descriptions leads to problems such as inaccurate image generation, lack of diversity in generated images, and unnatural outputs.

To address the aforementioned challenges, we have integrated CLIP (Contrastive Language-Image Pre-training) with Diffusion Models. CLIP, trained on a vast corpus of image-text pairs, has developed a profound understanding of cross-modal semantic relationships, demonstrating exceptional performance across a spectrum of visual and multimodal tasks. This integration is poised to tackle the three primary challenges outlined.

The main insight of this integrated approach lies in leveraging CLIP's robust text-to-image understanding capability to guide the diffusion model in generating high-quality images that more accurately align with the textual descriptions.

In summary, the integration of CLIP with Diffusion Models presents a formidable approach to overcoming the limitations inherent in Diffusion Models, offering a more efficient, statistically robust, and broadly applicable framework for text-to-image synthesis and beyond.

Related Work

Text-to-Image Synthesis with Diffusion Model

Diffusion Models are a class of deep generative models that generate new data samples by simulating the diffusion process of data. The core concept of these models is to gradually transform complex data distributions into simpler noise distributions and then reverse this process to generate new data samples. The working mechanism of Diffusion Models is primarily based on two mutually inverse processes: **the forward process** and **the reverse process**.

The forward process involves the model progressively introducing noise into the data until it is entirely converted into noise. This process is deterministic and can be precisely executed through exact mathematical calculations.

Conversely, the reverse process commences with noise and incrementally removes it to restore the original data. This process is typically learned through training and requires the approximation of deep neural networks to implement effectively.

These models have demonstrated a significant advantage in the quality of generated samples, particularly in image generation tasks, where they have shown the potential to outperform traditional Generative Adversarial Networks (GANs).

Stable Diffusion is a variant of the diffusion model that generates images by iteratively denoising data in the latent representation space, then decoding the results into complete images. This approach reduces the computational resources and time required in the text-to-image synthesis process while enhancing the quality and diversity of the generated images.

NoiseCollage is a layout-aware text-to-image diffusion model based on noise cropping and merging techniques. It generates multi-object images that reflect layout and text conditions, addressing the mismatch between text and layout conditions and the degradation of image quality during generation.

DiffAssemble is a unified graph-diffusion model for 2D and 3D reconstruction tasks. It treats elements of 2D patches or 3D object fragments as nodes in a spatial graph and reconstructs a consistent initial pose through iterative denoising. This model not only improves the quality of text-to-image synthesis but also enhances the model's adaptability to new domains and data types, thus enhancing the model's generalization capabilities.

However, methods solely based on diffusion models still face the three challenges mentioned above: semantic alignment between text and image, inconsistency in generated results and lack of diversity, and limited contextual understanding of the model.

CLIP

The CLIP (Contrastive Language-Image Pre-training) model is a multimodal pretraining neural network that employs contrastive learning to map images and text into a shared latent feature space, thereby enabling cross-modal

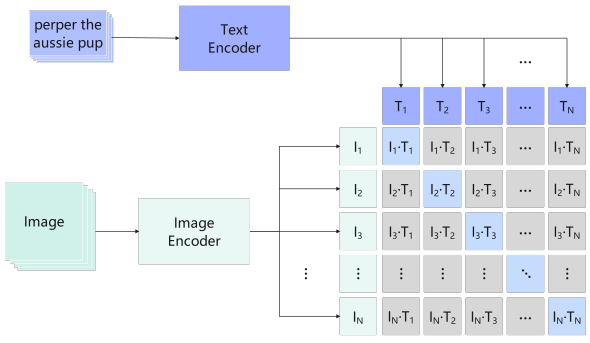

Figure 2: CLIP Decode

similarity matching. The core architecture of the CLIP model comprises two independent encoders: an image encoder for processing visual data and a text encoder for processing textual data. The image encoder typically utilizes

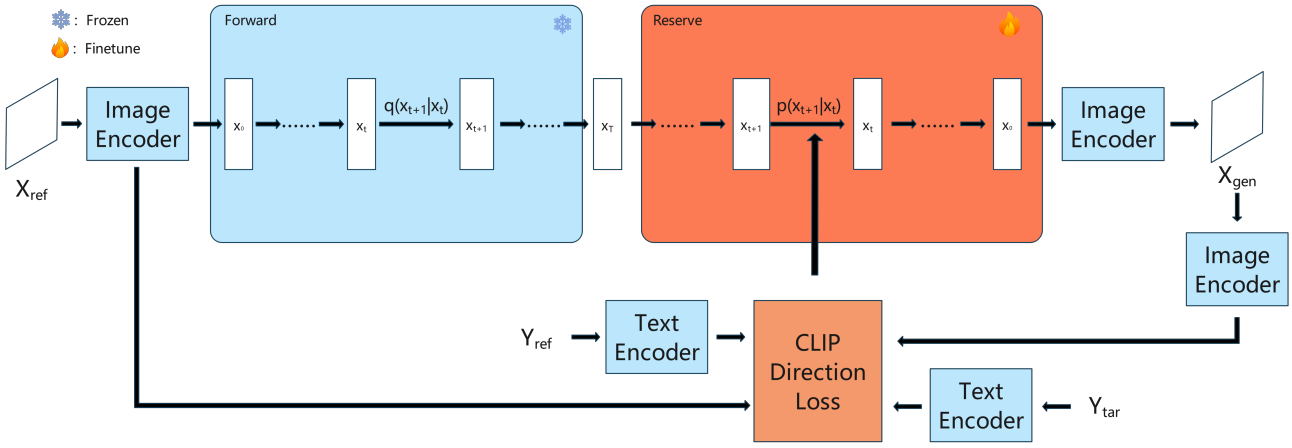

Figure 3: CLIP-Integrated Diffusion Model Framework

Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), while the text encoder adopts a Transformer architecture akin to the BERT model. This design enables the model to classify images directly through natural language descriptions without explicit labeling, demonstrating robust zero-shot learning capabilities.

In the realm of text-to-image synthesis, the CLIP model's primary advantage lies in its potent zero-shot learning ability. This means that even in the absence of explicit training samples for specific categories, the CLIP model can classify or generate new image or text samples based on the knowledge it has learned from large-scale datasets. The model's capacity for this kind of learning imparts broad application potential in the field of text-to-image synthesis.

The CLIP model's architecture and training process are conducive to its exceptional performance and generalization ability. Trained on a vast dataset comprising 400 million image-text pairs, CLIP learns rich associations between images and text through a weakly supervised pre-training approach. This enables the model to perform well in various downstream tasks such as image classification, image generation, and cross-modal retrieval, all without the need for task-specific fine-tuning. The simplicity of the CLIP model's architecture, coupled with its efficient training process, belies its remarkable performance, making it a prominent contender in the field of artificial intelligence for multimodal tasks.

Method

Workflow

First, the text undergoes pretraining, where the CLIP text encoder transforms the text into fixed-length vector sequences. These vector sequences encapsulate the semantic information of the text and are correlated with images from the real world. This allows the identification of which texts correspond to which images in the dataset, thereby establishing a bridge between text and image.

The next step involves the pretraining of the diffusion model. The pretraining of the diffusion model establishes an effective noise propagation and denoising relationship between the forward and reverse processes, enabling the model to generate realistic samples from noise. In the forward process, the model progressively adds noise to the original data until it is entirely transformed into random noise. This process is deterministic, with noise being incrementally added to the data at each step. The process can be expressed as follows:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t$$

In the reverse process, we introduce the CLIP directional loss, which will be specifically discussed in the next section. The reverse process begins with noise and progressively removes it to recover the original data. The entire process trains a neural network to approximate the inverse of the forward process. The process can be simply expressed as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t))$$

Throughout the sampling process of the diffusion model, we employ Denoising Diffusion Implicit Models (DDIM) for both the forward and reverse processes. DDIM represents a non-Markovian sampling method that significantly reduces the number of steps required in the reverse process while maintaining high generation quality. In this paper, the DDIM formulation can be rewritten as follows:

$$\frac{\sqrt{\frac{1}{\alpha_{t-1}}}x_{t-1} - \sqrt{\frac{1}{\alpha_t}}x_t}{\sqrt{\frac{1}{\alpha_{t-1}}-1} - \sqrt{\frac{1}{\alpha_t}-1})} = \varepsilon_\theta(x_t, t)$$

After completing the pretraining phase, as Figure **3** shows, the model can process input images and text. The input image undergoes the forward process of the pretrained diffusion model, where noise is progressively added to generate

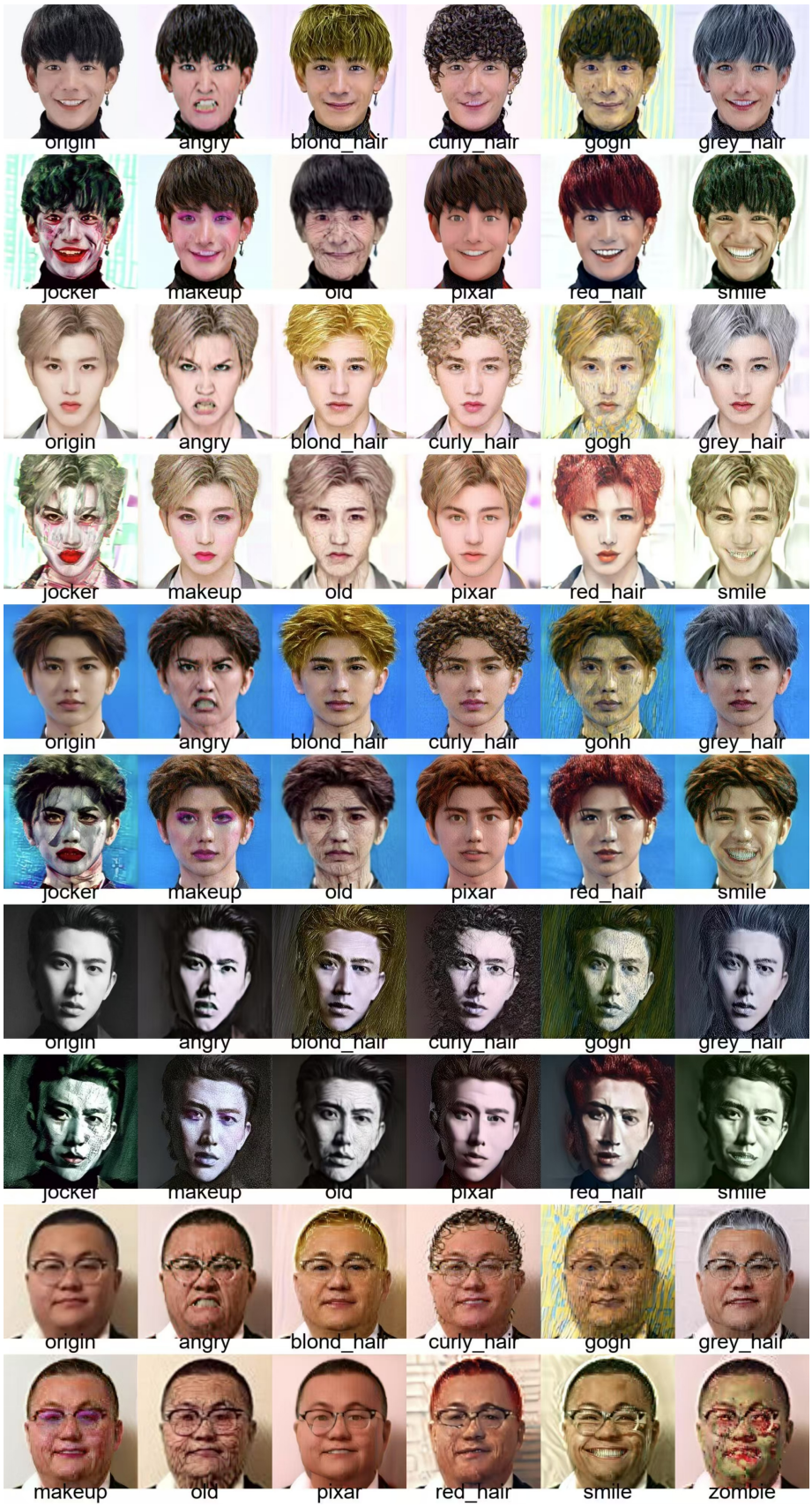

Figure 4: Generated Faces with CLIP-Integrated Diffusion Model

a noisy representation. Simultaneously, the input text is encoded using the pretrained CLIP text encoder. During the reverse process, the noisy image is iteratively refined based on the guidance of the CLIP text encoder, with fine-tuning performed under the supervision of the CLIP directional loss. Ultimately, this process produces a generated image that aligns with the semantics of the input text.

CLIP Directional Loss in Diffusion Model

Our model incorporates CLIP into the reverse path of the diffusion model by introducing a directional loss function defined by CLIP, which works in conjunction with the feature loss of the diffusion model. The input image is first processed through a pretrained diffusion model to generate latent variables. Guided by the integration of the CLIP directional loss function, the diffusion model is fine-tuned along the reverse path.The CLIP directional loss function can be expressed as follows (I: Image ; T: Text ; E: Encode):

$$1 - \frac{\langle [E_I(x_{gen}) - E_I(x_{ref})], [E_T(y_{tar}) - E_T(y_{ref})] \rangle}{\|E_I(x_{gen}) - E_I(x_{ref})\| \, \|E_T(y_{tar}) - E_T(y_{ref})\|}$$

The CLIP directional loss aligns the direction between the embeddings of the reference image and the generated image with the direction between the embeddings of a pair of reference text and target text in the CLIP space. The reason for selecting the CLIP directional loss function is that it helps mitigate the negative effects associated with the global CLIP loss, such as overfitting, susceptibility to adversarial attacks, low diversity, and the accumulation of semantic errors. By focusing on the directional relationship between the image and text embeddings, the model is better able to capture fine-grained semantic alignment, ensuring that the generated images are both more accurate and diverse. This approach not only improves the robustness of the model but also enhances its generalization ability, making it less prone to errors and more adaptable to a wider range of text-to-image synthesis tasks.

Experiments

In this section, we present the process of fine-tuning a pretrained model on input text to generate corresponding output images. The model has been primarily pretrained on facial data, with additional pretraining targeting specific types of textual inputs. The pretrained images are of 256×256 pixel resolution, which poses challenges in generating high-resolution, pixel-level images that meet expectations under CLIP guidance for specified text. This limitation could be mitigated by pretraining the model on higher-resolution images. Methodologically, our outputs also demonstrate the effectiveness of the proposed model.

To enhance the training of models tailored for facial data, we selected the CelebA-HQ dataset. For pretraining, we partitioned the dataset, using 70% of the images as training data and the remaining 30% as validation data.

During the pretraining process, we utilized two NVIDIA A100 GPUs, each equipped with 80GB of HBM memory. However, we did not specifically test the minimum GPU memory required for pretraining. Based on our observations, we recommend using GPUs with at least 16GB of memory. It is worth noting that most of NVIDIA's current mainstream GPUs, including server-grade models such as the NVIDIA A100 and NVIDIA H100, as well as PC-grade GPUs like the NVIDIA RTX 4060 and NVIDIA RTX 3060, support the Unified Memory mode. Therefore, in theory, as long as the combined memory capacity of the host system and GPU exceeds 16GB, the risk of encountering Out of Memory (OOM) errors during pretraining can be effectively mitigated.

It is important to note that during the pretraining process, the training dataset primarily consisted of images featuring clear, frontal portraits, where facial features were properly aligned. Consequently, when the model encounters portraits from side angles or unconventional perspectives, especially in conjunction with text inputs, it may produce unexpected or anomalous results. This limitation highlights the model's dependency on specific input conditions.

To ensure optimal performance, we emphasize the necessity of using frontal portraits, preferably with well-aligned and balanced facial features. That said, it is worth mentioning that as long as the facial structure is correctly positioned, the model can still generate reasonable outputs even if some facial features are partially missing. This reflects a degree of robustness in handling incomplete facial inputs but underscores the need for adherence to the dataset's primary characteristics to avoid unintended outcomes.

Figure **4** illustrates how input text prompts are utilized to perform semantic transformations on input images, aligning the visual output with the overall meaning conveyed by the text. These transformations are not limited to simple color changes, such as hair color or skin tone adjustments, but also extend to facial expressions, including smiles and other nuanced emotions. Furthermore, the approach supports more complex and stylistic modifications, such as applying makeup, creating a zombie-like appearance, or rendering the image in a two-dimensional (anime) style.

All five examples of input images demonstrate significant and perceptible changes, highlighting the versatility and adaptability of our method in achieving a wide range of text-guided facial transformations. This showcases the model's ability to handle both subtle adjustments and highly stylized alterations, ensuring alignment with the textual guidance provided.

Conclusion

In this paper, we integrate CLIP into the reverse process of a Diffusion Model, pretraining it on a facial dataset. Our proposed model enables users to input an image along with a specified textual description to perform image style transformation and generation. Experimental results demonstrate that the model achieves favorable performance in generating images of clear, frontal faces. However, its effectiveness is limited when dealing with faces that exhibit incomplete facial features or challenging angles.

References

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684-10695.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. 2022. Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data.In *Proceedings of the 40th International Conference on Machine Learning (PMLR)*. 4672-4712.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen,Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10696-10706.

Takahiro Shirakawa, and Seiichi Uchida. 2024. NoiseCollage: A Layout-Aware Text-to-Image Diffusion Model Based on Noise Cropping and Merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8921-8930.

Gianluca Scarpellini, Stefano Fiorini, Francesco Giuliari, Pietro Moreiro, and Alessio Del Bue. 2024. DiffAssemble: A Unified Graph-Diffusion Model for 2D and 3D Reassembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 28098-28108.

*Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka.*Sean: Image synthesis with semantic region-adaptive normalization. 2020. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 5104-5113.

Alex Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *arXiv preprint*. arXiv:2102.09672.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan,and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of International Conference on Machine Learning (PMLR).* 2256-2265.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 8188-8197.

G. Kim, T. Kwon and J. C. Ye. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation.In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2416-2425.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684-10695.

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. Stylegan-nada: Clip-guided domain adaptation of image generators. In *arXiv preprint*. arXiv:2108.00946.

Clay Mullis and Katherine Crowson. Clip-guided diffusion github repository. In *https://github.com/afiaka87/clip-guided-diffusion.*

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *arXiv preprint*. arXiv:2103.17249.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 4690-4699.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis. In *arXiv preprint*. arXiv:2105.05233.

Tero Karras, Samuli Laine, and Timo Aila. 2017. Progressive growing of gans for improved quality, stability, and variation. In *arXiv preprint*. arXiv:1710.10196.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2256-2265.