Ideal Face Generation Model

Kyutae Kim
Data Sciene
Sungkyunkwan University
rlarbxo0324@gmail.com

Sunwoo Hwang
Artificial Intelligence
Sungkyunkwan University
sunwoo1357@g.skku.edu

Sohyun Lee
Artificial Intelligence
Sungkyunkwan University
wodosl902@gmail.com

Yeoeun Yu
Artificial Intelligence
Sungkyunkwan University
annie2675@g.skku.edu

Abstract

This study aims to develop a ideal face image generation by text description. We trained a VQGAN model using the AI-Hub montage dataset and integrated this model as the encoder and decoder for the KoDALLE framework. Subsequently, we refined the output of the KoDALLE model to align with aesthetically pleasing facial features using StyleGANEX, which had been trained on CelebA dataset. This approach aimed to generate final outputs that closely resemble idealized human appearances. Despite the seamless progression during the training phase, the process was hindered by limited computational resources, preventing sufficient model training. Consequently, the validation results were suboptimal, and the final application did not produce satisfactory outcomes.

1. Introduction

In modern society, an individual's dreammate holds significant importance to many people. Technologies that can concretize the appearance characteristics of a dreammate and visually represent them can positively impact personal psychological satisfaction and social interactions. However, existing technologies that describe the appearance of a dreammate and generate a corresponding face are limited. People often imagine their dreammate in their minds but struggle to materialize this image. To address this issue, there is a growing need for a service that generates the face of a dreammate based on appearance descriptions. This service can provide a personalized experience by visually expressing an individual's dreammate, thereby enhancing psychological satisfaction and improving interactions between users on social networks and dating apps. Additionally, the technology of generating dreammate faces contributes to psychological research on human ideals, allowing for a better understanding of the psychological significance and impact of a dreammate. Furthermore, the industrial applications of this technology are vast, including customized advertising, virtual character creation, and production in the film and gaming industries. For these reasons, a model framework that generates a dreammate's face based on appearance descriptions is a highly significant research topic, warranting continuous attention and investment. Recently, with the development of generative models, some generative models for face generation have emerged, making such service possible. Additionally, models for face manipulation are also developed. In this study, we build a pipeline with a multimodal model KoDALLE [8] and a manipulation model StyleGANEX [6].

2. Related works

Face genterative models. To generate facial images, previous researches have generated facial images through various methods, such as generating several facial elements and combining them or generating a face with text describing the appearance at once. MontageGAN [5] leverages a two-step GAN framework: local GANs generate specific facial parts in separate layers, while a global GAN assembles these layers into a complete facial image. Meanwhile, a research on Generative AI for Korean Multi-modal Montage App [8] developed an app that takes a facial description in Korean as input and outputs a montage image. This app utilizes KoDALLE, which consists of KLUE RoBERTa-large and VQGAN. VQGAN [2] is a model that combines the advantages of CNN and Transformer, which is learned through advertising learning and constructs images using Transformer. KoDALLE is based on DALL-E [3], which is a model that trains transformers by accepting text and image tokens as a stream.

Face manipulation using StyleGAN. In addition to the re-

searches on generating facial images, researches on modifying a style of images have also been conducted. The StyleGAN is mainly used for such researches. For instance, to create more natural virtual dreammate images that a user might prefer, photographs of celebrities chosen by a user are synthesized using StyleGAN [7]. And [6] designed StyleGANEX, an enhanced version of StyleGAN, to improve its performance and enable additional style manipulations such as facial attribute editing and sketch-to-face translation.

3. Data

We used 페르소나 기반의 가상 인물 몽타주 데이터 on the AI-Hub to train the VQGAN model. This dataset is based on personas and contains montage images of fictional characters. It includes detailed textual data about facial features such as eyes, nose, mouth, and eyebrows for each persona. This level of detail provides a foundation for the KoDALLE model to generate similar montages based on the input text. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=618

4. Model

We used three models to generate a ideal face based on input text content. Each three models are used for image generate training, text layer connect, and image style transform.

4.1. VQGAN

VQGAN(Vector Quantized Generative Adversarial Network) is image generating model which combined Vector Quantized Variational Autoencoder (VQ-VAE) and Generative Adversarial Network (GAN) to generate high-resolution images.[2] We trained 페르소나 기반의 가상 인물 몽타주 데이터 to this model.

4.2. KoDALLE

KoDALLE It is a Korean-based image generation model that replaces GPT-3 with the KLUE RoBERTa-large model in the DALL-E structure.[4] This model is capable of taking Korean text as input and producing visual representations of the described content. We utilized an extracted RoBERTa model for the encoder component and employed a trained VQGAN model for the decoder component.

4.3. styleGANEX

StyleGANEX is an advanced generative model derived from the StyleGAN architecture, specifically designed to produce highly realistic and high-resolution images. This model extends the original capabilities of StyleGAN by incorporating various enhancements that improve its versatility and effectiveness.

5. Experiment

In this project, our team used aiHub's 'Persona-based Virtual Person Montage Data' and trained VQGAN with an image to use VQGAN as an encoder and decoder. Then we trained KoDALLE with label and image data to generate a montage image corresponding to the input text.

5.1. Preprocessing

To preprocess the prepared dataset, JSON files located in the dataset and containing information of each montage image were used. In particular, 'description' values which describe the impression of montage images and 'img_path' values which contain paths of montage images that those descriptions refer to were used. We extracted all montage datas according to the 'img_path' values and saved them in a newly generated 'images' folder. We also saved the 'description' values in a newly generated 'labels' folder. Lastly, we created text files containing the paths of newly saved images and labels.

5.2. Training

5.2.1 VQGAN

VQGAN model was trained on a the Google Colab utilizing an L4 GPU. To train the model, we cloned the taming-transformer [1] repository from GitHub to leverage its modules and load pre-trained VQGAN model. The input images from the dataset were transformed to a size of 256x256 pixels. The dataloader was configured with a batch size of 8. the Negative Log Likelihood (NLL) loss function was used to guide the training. Given that each epoch taking approximately 6 hours, an encoder and a decoder to learn the feature of the painting style of the montage image was trained with 8 epoch.

5.2.2 KoDALLE

KoDALLE model training was carried out using a GPU in the Google Colab T4 environment. The model was cloned from the Git-hub and the 'vae-config' file was constructed using pretrained VQGAN's ckpt file. We extraced huggingface's robusta-large model and used as a text embedding layer, and an image patch was created through 16 layers of the image patch generator using this text embedding layer. After that, an image is generated using the VQGAN Decoder, an image generator. We used Weight and Bias' platform to measure the performance of each model generated as a result of the learning.

5.3. Style Transform

The generated face was put into StyleGANEX, which was pretrained by FFHQ Dataset. Inversion and style mixing

were carried out trough the model. The model first transformed the generated face into a w^+ vector so that it could be manipulated. Random vectors following the form of the w^+ vector were generated for the change of face style. Those vectors were replaced with the original w^+ vector in the process of reconstruction. Through this process, various styles of faces were able to be generated from one image.

6. Result

Through the above experimental process, we successfully generated images based on text and performed style transfer. The sequence of steps includes training the VQGAN, which acts as the image encoder and decoder in KoDALLE, generating images using KoDALLE, and performing style transfer using StyleGANEX. The detailed explanation of the process is as follows:

6.1. Encoder/Decoder Training

The VQGAN was trained to function as both the image encoder and decoder within the KoDALLE framework. This step involved training the model on a large dataset to ensure it could accurately encode images into discrete latent codes and subsequently decode these codes back into images.

Figure 1. Input

Figure 2. Reconstruction

6.2. Image Generation

Utilizing the trained VQGAN, we employed KoDALLE to generate images from textual descriptions. This involved feeding text inputs into the model, which then produced corresponding images based on the learned associations between text and visual content. Input text was ”바가지 모양의 헤어 스타일이 촌스럽게 보이지만 내려간 눈꼬리와 살짝 올라간 입꼬리는 선한 이미지로 보여진다. 중간 톤의 피부에 주름 없고 깔끔하게 정리된 수염은 건강하고 성실해 보이며 믿음이 가는 인상이다. [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]”. Fig.3 is the output of the KoDALLE based on input above.

6.3. Style Transfer

After generating the initial images with KoDALLE, we applied StyleGANEX for style transfer. This step involved taking the generated images and transforming their style to match a desired aesthetic, leveraging the capabilities of StyleGANEX to manipulate the visual attributes while preserving the underlying content. Fig.4 is the picture which is

Figure 3. Image Generated by KoDALLE

inversed by styleGANEX from Fig.3. Through this process, we were able to soften the potentially harsh atmosphere of the montage and transform it into a more aesthetically pleasing facial image. Subsequently, by employing the style mixing process, we generated a variety of different atmospheres for the same face like 5. This approach enhances the likelihood of users finding a preferred style for the depicted individual.

Figure 4. Image Inversion

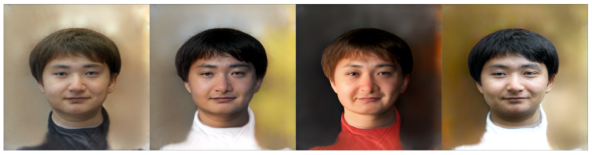

Figure 5. Style mixed image

7. Discussion

The experimental results demonstrated success in generating images based on text information, transforming them into aesthetically pleasing images, and creating various atmospheric facial images through style mixing. This experiment provides a method to concretize the ideal image that people may not have had a clear picture of in their minds.

Limitation of this study was the substantial resources required to train on a vast number of facial images and their

corresponding descriptions. Due to resource constraints, a compromise had to be made. In future research, we aim to address this limitation by securing more resources, allowing for higher quality results and more comprehensive experiments.

References

[1] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 2

[2] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. 1, 2

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1

[4] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[5] Chean Fei Shee and Seiichi Uchida. Montagegan: Generation and assembly of multiple components by gans, 2022. 1

[6] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Styleganex: Stylegan-based manipulation beyond cropped aligned faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21000–21010, 2023. 1, 2

[7] 임연수, 김창민, 석정민, 박덕원, and 김태형. Stylegan 을 이용한 이상형 생성 웹 서비스 개발. *한국통신학회 학술대회 논문집*, pages 1368–1369, 2020. 2

[8] 임정현, 차경애, 고재필, and 홍원기. 한국형 멀티모달 몽타주 앱을 위한 생성형 ai 연구. *서비스 연구*, 14(01):13–26, 2024. 1