S1 Signature Work Report

**Title**: A Research on Merging Human Image and Text Image Using New Algorithm

**Acknowledgement:**

Thanks to Prof. Luyao Zhang for her great help in this project. She provided me with valuable professional suggestions.

**Disclaimer:**

This article is a final deliverable from Jiaolun Zhou’s signature work entitled “**A Research on Merging Human Image and Text Image Using New Algorithm**,” guided by Prof. Luyao Zhang at Duke Kunshan University.

**How I apply what I have learnt from the courses and what the differences are:**

During my four years of research at the university, I learned a lot about graph neural networks and image-text fusion under the guidance of my supervisor, Luyao Zhang, which was also my main research direction during my four years of university. I also learned how to use models to solve social problems and the process of doing a professional research by participating in Prof. Luyao Zhang’s Econ211 “Intelligent Economics: An Explainable AI approach” instructed by Prof. Luyao Zhang, Econ206 “Computational Microeconomics” instructed by Prof. Luyao Zhang, and other disciplines integrating computers and economics, which provided guidance on research methodology for my signature work. Before that, I conducted a research named “FaceCDM: Any Face Can Be Inserted into Generated Images through Diffusion Model”, which focused on using graph neural networks to generate fixed face images, and my signature work is also related to this aspect.

In my signature work, I used different techniques in image text fusion, mainly including face feature generation and modification techniques based on the implementation of multi-layer deep learning systems, the application of generative adversarial networks in face image generation, image generation methods based on specific text and face features, and multi-layer semantic information separation methods based on StyleGAN2 model. All of this knowledge comes from studying and summarizing previous papers in my previous research, and the same is also inspired by experimental learning.

**Abstract**:

This research proposes to develop a new algorithm that can combine specific text and face features to generate new images. Although generative adversarial networks have been very successful in face image generation and editing, it is still a major challenge to manipulate facial semantic attributes in the potential coding space. In order to overcome the shortage of labeled data, some recent works have attempted to accomplish the above task with the help of pre-trained models. However, the accuracy of this approach and the fidelity of the results do not fulfill the needs of real face editing scenarios. Therefore, this thesis delves into the algorithm for fusing specific text and face features to generate new images, and proposes a novel algorithm that can effectively fuse text descriptions with face features to generate high-quality, personalized new face images. At the level of experimental results, this study intends to compare whether the algorithm has significant improvement in subjective visualization and compare its performance on PSNR and SSIM with existing methods.

**Research Question**:

1. How to combine machine learning and human computer dialogues to transform text into actual images?

**Methodology:**

The method employs the preprocessed CelebA dataset for training and testing, which can ensure that the generated images have high quality and diversity. In addition, the method employs a generative adversarial network framework based on the attention mechanism, which can effectively extract and retain important features in face images, thus improving the quality and diversity of the generated images.

**Dataset:**

CelebA is a dataset containing over 200,000 images of human faces, each labeled with 40 attributes (e.g., age, gender, expression, etc.). This dataset is widely used in the fields of face recognition, face keypoint detection, and face attribute analysis. In my research, I will use the preprocessed CelebA dataset to verify the effectiveness of their algorithm. Since the CelebA dataset contains rich facial feature information, it is well suited for investigating how to generate high-quality face images by combining textual descriptions with facial features. At the same time, the CelebA dataset is also publicly available, facilitating replication and further research by other researchers.

**Intellectual Merit:**

The intellectual merit of this research lies in introducing cutting-edge techniques and research directions, such as combining machine learning with human-computer dialogues to transform text into images, using multi-layer deep learning systems for facial feature generation and modification, and exploring the application of Generative Adversarial Networks (GANs) in facial image generation. These advancements offer new ideas and methods for face image generation and are expected to find applications in virtual reality, gaming, and other fields.

**What did you do?**

In my signature work, I explore cutting-edge technologies and research directions in face image generation. This includes converting text into images using machine learning and human-computer dialog, implementing face feature generation and modification with multi-layer deep learning systems, and applying generative adversarial networks (GANs) for face image generation. I also delve into specific techniques, such as combining text and face features to generate images and using a multi-layer semantic information separation method based on the StyleGAN2 model. Overall, my work provides new ideas and methods for face image generation.

**What did you produce?**

My research focuses on the challenges in face image generation and editing, particularly in manipulating semantic attributes within the underlying coding space. While Generative Adversarial Networks (GANs) have made significant strides in face image generation, they still face difficulties in handling semantic attributes for face editing. This is largely due to the need for extensive labeled data, which is both costly and time-consuming to collect and label.

To address the issue of limited labeled data, some studies have explored using pre-trained models. Although this approach can achieve reasonable results, it often falls short in terms of accuracy and realism, especially in practical face editing scenarios. Additionally, using GANs for image generation can lead to issues such as low resolution, blurred edges, and loss of identity information.

My research proposes a new algorithm that effectively fuses specific text descriptions with face features by analyzing the correlation between text descriptions and face features in order to generate high-quality and personalized new face images. Experiments using the preprocessed CelebA dataset show that the algorithm provides a significant improvement in subjective visualization as well as improvements in objective metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM). This indicates that the algorithm has improved both the accuracy and diversity of image generation.

Overall, my research lays a good foundation for further research and application of the technique of fusing text and face features to generate new face images, and is of great significance in advancing the field of artificial intelligence.

**Significance:**

The significance of this research in the current literature is that it not only improves the user experience by combining machine learning and human-computer dialogue systems to allow users to express their needs through natural language or image input, making the image generation process more intuitive and easy to use, lowering the technological threshold, and allowing more non-professionals to enjoy high-quality image generation services; at the same time, it advances the development of cross-modal transformation technology, and studying the cross-modal conversion problems involved in the process of converting text descriptions into images helps to better understand the relationship between different forms of information and promotes the development of multimodal information processing technology; in addition, the new infrastructure constructed using the optimized StyleGAN2 and CLIP algorithms can improve some of the shortcomings of the existing methods, such as the realism of the generated image insufficiency, feature matching failure, etc., so as to improve the overall quality and clarity of the generated images; more importantly, it solves the problem of semantic consistency, which is one of the key challenges to ensure the consistency between the image content and the textual description during the text-based facial image generation process, and by investigating how to find the effective connection points to match the text embedding vectors with the visual attributes in the generative adversarial network, it the problem of image-text semantic unification can be further addressed. I believe that my research contributes to the advancement of large modeling at the theoretical level, and can provide a new way of thinking and promote the development of modeling in the field of image-text fusion.

**How this project is helpful to my personal development:**

As a student majoring in Computer Science, this research enables me to gain an in-depth understanding of deep learning models, especially Generative Adversarial Networks (GANs) and their variants like StyleGAN, along with the working principles and optimization techniques of these models. Through the study of these models, I can master how to design and train complex neural networks to generate high-fidelity images.

Secondly, this study emphasizes the cross-integration between Natural Language Processing (NLP) and Computer Vision (CV), which is of vital significance for understanding and developing advanced cross-modal systems. I can thereby learn how to effectively parse natural language and transform it into structured information that can be utilized to guide image generation. This is highly beneficial for future research in domains such as speech recognition, automatic translation, or intelligent chatbots.

Furthermore, the research also focuses on enhancing the authenticity and consistency of the generated images. This involves the design of image encoders and decoders, the techniques of feature extraction and matching, as well as the establishment of standards for evaluating the quality of the generated images. These are core issues in computer vision and can assist me in establishing professional skills in image synthesis, editing, and analysis.

Finally, this study showcases how to overcome issues of training instability and low output quality through optimization algorithms and architectures. This is an extremely valuable experience for any student who aspires to make in-depth advancements in the field of artificial intelligence. It teaches me how to confront and solve complex challenges in practical applications, thereby enhancing the overall work efficiency and quality of the outcomes.

**Conclusion**

In the process of creating my signature work, I have gained invaluable experience and significant growth in my technical skills. This project has deepened my understanding of cutting-edge technologies in face image generation and editing, particularly in the areas of machine learning, deep learning, and generative adversarial networks (GANs). The challenges I encountered and overcame have honed my problem-solving abilities and reinforced the importance of innovation in addressing complex issues. Overall, this work has not only enhanced my expertise in computer science but also inspired me to continue exploring and advancing in this exciting field.

