FETTER: FACIAL EMOTION AND TEXTURE TRANSFER WITH EFFICIENT REPRESENTATIONS

AHMET YAYLALIOĞLU

Thesis for the Master's Program in Electrical and Electronics Engineering

Graduate School
Izmir University of Economics
Izmir
2024

FETTER: FACIAL EMOTION AND TEXTURE TRANSFER WITH EFFICIENT REPRESENTATIONS

AHMET YAYLALIOĞLU

THESIS ADVISOR: ASSOC. PROF. DR. MEHMET TÜRKAN

Master's Exam Jury Members
Prof. Dr. Devrim ÜNAY
Assoc. Prof. Dr. Pınar OĞUZ EKİM
Assoc. Prof. Dr. Mehmet TÜRKAN

A Master's Thesis
Submitted to
the Graduate School of Izmir University of Economics
the Department of Electrical and Electronics Engineering

Izmir
2024

ETHICAL DECLARATION

I hereby declare that I am the sole author of this thesis and that I have conducted my work in accordance with academic rules and ethical behaviour at every stage from the planning of the thesis to its defence. I confirm that I have cited all ideas, information and findings that are not specific to my study, as required by the code of ethical behaviour, and that all statements not cited are my own.

Name, Surname:

Ahmet YAYLALIOĞLU

Date:

01.07.2024

Signature:

ABSTRACT

FETTER: FACIAL EMOTION AND TEXTURE TRANSFER WITH EFFICIENT REPRESENTATIONS

Yaylalıoğlu, Ahmet

Master's Program in Electrical and Electronics Engineering

Advisor: Assoc. Prof. Dr. Mehmet Türkan

June, 2024

In the current era, advancements in generative networks, alongside hardware with high computational capacity, have made the outputs of artificial intelligence research indistinguishable from reality. Numerous Generative Adversarial Network (GAN) models for synthesizing and manipulating human face images exist. However, the high output quality of generative networks requires substantial computational power, necessitates working with fixed-sized images, and demands large datasets containing vast amounts of data. Additionally, these models inherit the characteristics of the datasets on which they are trained. Our novel approach demonstrates that texture and emotion transfers can be quickly performed between human face images of any size using Discrete Cosine Transform (DCT) dictionary based sparse representation

and gradient-descent optimization. Our method, which incorporates image pyramids and facilitates rapid optimization at every pyramid level, has produced stable results comparable to state-of-the-art GAN outputs.

Keywords: facial texture transfer, sparse representations, gradient descent optimization, image fusion .

ÖZET

FETTER: ETKİLİ TEMSİLLER İLE İNSAN YÜZ GÖRÜNTÜLERİ ARASINDA DUYGU VE DOKU AKTARIMI

Yaylalıoğlu, Ahmet

Elektrik-Elektronik Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Mehmet Türkan

Haziran, 2024

Günümüzde, yüksek hesaplama kapasiteli donanımlarla birlikte gelişen üretici ağlar, yapay zeka araştırmalarının çıktılarını gerçekten ayırt edilemez hale getirmiştir. İnsan yüz resimlerini sentezlemek ve manipüle etmek için birçok Çekişmeli Üretici Ağ (GAN) modelleri bulunmaktadır. Ancak, üretici ağların yüksek çıktı kalitesi, önemli hesaplama gücü gerektirir, sabit boyutlu resimlerle çalışmayı zorunlu kılar ve geniş veri miktarları içeren büyük veri kümeleri talep eder. Ayrıca, bu modeller eğitildikleri veri kümelerinin özelliklerini miras alır. Sunduğumuz yenilikçi yaklaşım, Ayrık Kosinüs Dönüşümü (DCT) sözlüğü tabanlı seyrek temsiller ve gradyan iniş optimizasyonu kullanarak herhangi bir boyuttaki insan yüz resimleri arasında doku ve duygu aktarımlarının hızlı bir şekilde gerçekleştirilebileceğini göstermektedir.

Görüntü piramitlerini içeren ve her piramit seviyesinde hızlı optimizasyonu sağlayan metodumuz, modern GAN çıktıları ile karşılaştırılabilir kararlı sonuçlar üretmiştir.

Anahtar Kelimeler: insan yüzlerinde görsel doku transferi, seyrek temsilleme, gradyan iniş optimizasyonu, görüntü füzyonu.

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to my advisor, Assoc. Prof. Dr. Mehmet Türkan. His exceptional guidance throughout my master's degree, his ability to reignite my motivation during challenging times, and his career advice have been invaluable. His patience, dedication, and willingness to assist are truly commendable. The opportunity to work with him in such a supportive and friendly environment has been a privilege.

I extend my thanks to the jury members, Assoc. Prof. Dr. Pınar OĞUZ EKİM and Prof. Dr. Devrim ÜNAY, for agreeing to be part of my thesis defense. Their insightful feedback and constructive comments significantly enhanced the quality of my work.

Furthermore, I want to express my deepest and most heartfelt thanks to my beloved Semanur Uç, who has always stood by me, even under the most challenging circumstances. Semanur has supported me not only with my thesis but also with everything I needed to overcome. I feel very fortunate to have her in my life. I would also like to thank my greatest source of motivation, my mother, Gülseren Yaylalıoğlu, my father, Osman Yaylalıoğlu and, my sisters, Hande and Zeynep Yaylalıoğlu, for all their support and trust in me. Lastly, I thank my colleague, Furkan Karagöz, for all his support and valuable comments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In human face images, transferring the texture and emotional expressions between faces is a necessary and intriguing application area in the digital world. The production of high-computing power hardware and advancements in generative neural networks have brought intelligent visual editing applications in sectors like entertainment, the film industry, and photo montage. Moreover, the presence of camera equipment in many places today has raised concerns about the privacy of visual data. Thus, texture transfer between faces can also be utilized to anonymize face images. Additionally, in the medical aesthetics field for pre-procedure image alterations, in forensic investigations for generating different versions of existing photos (like aging), and in computer vision and deep learning research to enhance dataset diversity in a consistent and noise-free manner, the concept of texture and emotion transfer in face images can be beneficial.

In the past decade, particularly with the increase in dataset diversity and the rise in computing power of graphics card hardware, robust working algorithms, and model architectures have been developed in the field of generative artificial intelligence. From Generative Adversarial Network (GAN) (Goodfellow et al., 2014) models that produce handwritten numeric characters consisting of binary black-and-white pixels to GAN and Diffusion-based generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020) producing human face images indistinguishable from real humans, a wide range of generative model architectures (Radford et al., 2015; Zhu et al., 2017; Isola et al., 2016; Karras et al., 2018) have been published. With the development of multi-modal embedding models (Ma et al., 2024), generative vision models combined with large language models can now produce unique new images and edit existing images based on natural language inputs (Ramesh et al., 2021; Rombach et al., 2021). In these generative AI models, since the probability distribution function of images and pixels in the image dataset is predicted conditionally or unconditionally, the images generated by sampling this probability distribution with random noise are unique. Due to the models' capacity for unique image production, generative AI models are frequently

utilized for digital content creation, fake content production, or identity concealment. Despite producing highly successful outputs, these generative AI models also have particular challenges. The most significant challenges include high hardware power, massive datasets, unwanted blurriness and distortions in some areas of the output images, and changes in the identity of the person whose facial features are transferred (Wu et al., 2020). Additionally, these large models produce low-resolution/small-sized images in the output since they operate with a fixed input size, which can vary depending on hardware power. Furthermore, since generative network models reflect the characteristics of the dataset they were trained on, undesirable distortions can occur in the output image when they receive an input that is very different from the dataset collection.

Besides generative AI models, there are methods based on Convolutional Neural Networks (CNN) (Gatys et al., 2016; Kaur et al., 2017), Dictionary Learning (Abiantun et al., 2019; Huang et al., 2010; Yang et al., 2016), 3D Morphable Face (Thies et al., 2016; Garrido et al., 2014), and probabilistic modeling (Frigo et al., 2016). CNN-based methods are mainly used for neural style transfer (Gatys et al., 2016) and have produced successful results. In the neural style transfer method, the necessary weights are obtained using a CNN-based model and applied to the input image, creating an impression-like human face crafted with an artist's brush strokes. Additionally, neural style transfer has been performed using classical approaches, where the most suitable patches in the reference image are combined most appropriately with the input image pixels considered as a graph using probabilistic graphical models. However, these style transfer approaches do not fully align with the method presented in this article, as the image style is transferred across the entire image. In the presented method, large and small wrinkles around the human face and eye area, along with the mouth structure of the reference image, are seamlessly transferred or fused into the input image.

Furthermore, some methods based on Dictionary Learning and Sparse Representation successfully enhance the resolution of face images, inpainting missing parts of faces in the most suitable way within a trained dictionary (Abiantun et al., 2019) and aging faces (Huang et al., 2010). However, compared to the method presented in this article, using a fixed Discrete Cosine Transform-based dictionary eliminates

dependence on a training process and dataset, presenting a more general approach. Nevertheless, since the presented method is unsuitable for real-time operation, it cannot be compared with methods utilizing computer graphics discipline based on 3D Morphable Face for real-time facial reenactment.

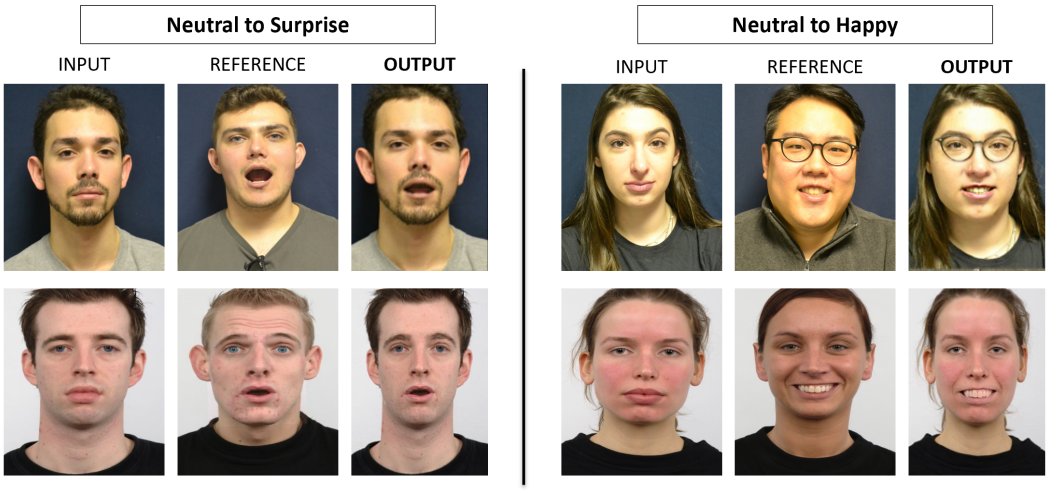

Figure 1. Some result of proposed method,first row: neutral to surprise and happy examples from TUFTs Emotion Dataset, second row: neutral to surprise and happy examples from RafD Dataset

The method presented in the article offers practical benefits. It produces results with a general approach independent of a dataset, which, unlike generative AI models, is more inclined to produce outputs free from noise and hallucination. Its unique feature of operating solely on an optimization basis, without requiring a dataset, allows it to run on lower-capacity hardware and produce more stable results. Moreover, its independence of the input image size and direct applicability to all human face images are distinguishing features compared to generative AI models.

In this paper, we present our method using sparse representation and gradient-descent-based optimization without any dataset. We use two human face images to transfer the texture and emotional state from one to the other, and we describe the outputs we obtained. In our method, we opted for a fixed Discrete Cosine Transform-based dictionary and Orthogonal Matching Pursuit (OMP) (Cai and Wang, 2011) in our sparse representation approach. According to our dictionary, this approach will work successfully on all human face images with a global approach without retraining the

dictionary. Our method used the sparse representation display as a feature extractor. The features in the human face images we wanted to transfer (mouth, wrinkles, pits, and protrusions) were associated with the term texture, so we approached the problem as a Cartoon - Texture Separation (Elad et al., 2005) concept. We matched the skin tones of the input and reference images according to the histogram of the input image. We warped them onto each other using Delaunay Triangulation (Lawson, 1972) with facial landmarks. Subsequently, we minimized the designed loss function with the texture component of the image obtained from sparse representation using the gradient-descent algorithm. We used image pyramids to optimize the fine details quickly and effectively. This way, we can apply our solution to images of any size without being restricted by a specific resolution constraint.

Figure 1 presents several successful results of the method explained in the following sections of this article. This method alters the emotional state and texture of the input human face to match the reference image while maintaining the face's identity and eliminating signal noise.

This paper is organized as follows. Section 2 explains similar approaches found in the literature while highlighting the differences of the presented method. Section 3 explains the proposed optimization based facial texture and emotion transfer approach in detail. Section 4 details the experimental setup and greatly discusses the obtained results in comparison to the competing algorithms in the literature. Lastly, Sec. 5 provides a brief conclusion for this study.

CHAPTER 2: LITERATURE REVIEW

Advances in machine learning have significantly advanced facial expression manipulation and texture transfer methodologies. These advancements, which include deep generative models, disentangled representations, and image fusion techniques, have practical applications in various fields. This section delves into these methodologies, with a particular focus on their facial editing and expression transfer applications.

2.1. Optimization and Sparse Representation Based Methods

Image fusion based on sparse representation methods has been producing successful outputs. In environments captured with two different focus settings without changing the camera position, focused areas of the two images can be fused into a single image. In Yang and Li (2012), the Orthogonal Matching Pursuit method uses a fixed dictionary basis to combine pixels from the focused regions of two images. A series of filters was used to detect focused regions. The method in the Yang and Li (2012) is applied based on the entire image. In contrast, the method presented in this paper, although tailored for human face images, serves as a fine example of low-level image fusion work in pieces. Another publication utilizing sparse coding, Yang et al. (2016) uses a specific training set to decompose human face images into average face, identity, age, and noise components with the aid of Hidden Factor Analysis (Yang et al., 2016) and then sparsely represents these components through sparse coding. The method in Yang et al. (2016) requires a dataset containing images of elderly individuals. Wrinkles and pits on the sparsely represented face are merged in the input image to visually demonstrate what an older version of the person might look like. The method presented in this paper optimally transfers the wrinkles and pits from the reference image to the input image as a texture without depending on a dataset. In another study of face aging using sparse representation, Huang et al. (2010) identifies the most suitable vectors using Maximum A Posteriori (MAP) from image vectors represented sparsely using a small dataset. After an affine transformation for geometric transformation between faces, texture synthesis is performed using Markov Random

Field. Another area where sparse representation is effective is the single-image super-resolution. In this field, the SSR2 method (Abiantun et al., 2019) competes with Generative Adversarial Network models and not only enhances extreme low-resolution images to high resolution but also optimally restores parts of a human face image that might be missing using the trained dictionary. In the SSR2 method (Abiantun et al., 2019), a dataset of high-resolution human face images is trained using image pyramids at each level of the pyramid with the dictionary learning method. When a low-resolution face image is input into the system, the method matches it to the scale level of an image pyramid based on resolution and completes the missing parts using the dictionary learned at other levels of the pyramid. Therefore, even if certain parts of a human face image are missing, the gaps can be filled using a dictionary. Another approach, Benning et al. (2017) uses Total Variation Regularization (TV) to decompose human face textures (wrinkles, pits, etc.) according to their levels and transfer them to a new image. This decomposition allows for the fusion of specific features across images, a crucial task in blending textures or expressions from different faces. Such methods provide a basis for more nuanced and controlled manipulations. In our comparisons, identity loss has been detected with transfers made using this method. The method we present performs the image fusion at a more low-level, producing outputs that are closer to reality and more prone to preserving identity. Deep learning and generative network-based approaches can also be examined under the following headings.

2.2. Generative Adversarial Networks and Deep Learning Based Methods

Facial editing techniques like STD-GAN (Guo et al., 2021) and Cascade EF-GAN (Wu et al., 2020) utilize GANs to achieve realistic, high-fidelity manipulations. STD-GAN integrates style and texture discriminators to refine the quality and accuracy of attribute transformations. At the same time, Cascade EF-GAN employs a multi-stage approach to adjust significant facial features locally, enhance expression clarity, and reduce visual artifacts. These models illustrate the effectiveness of GANs in producing detailed and context-aware facial attribute modifications. StarGAN (Choi et al., 2018), on the other hand, can manipulate multiple facial attributes within a single model by

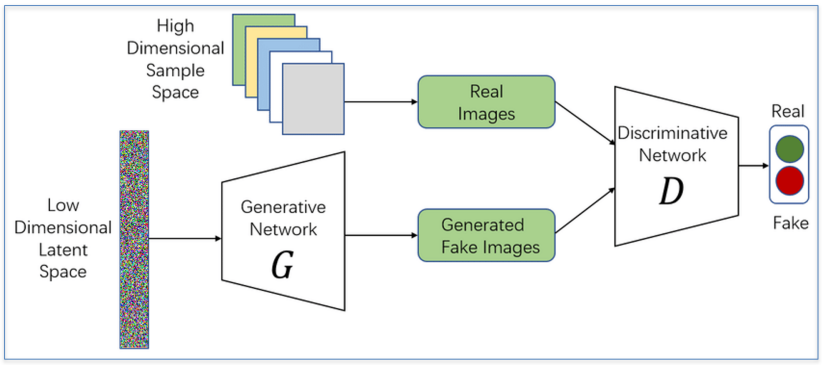

Figure 2. Basic GAN architecture

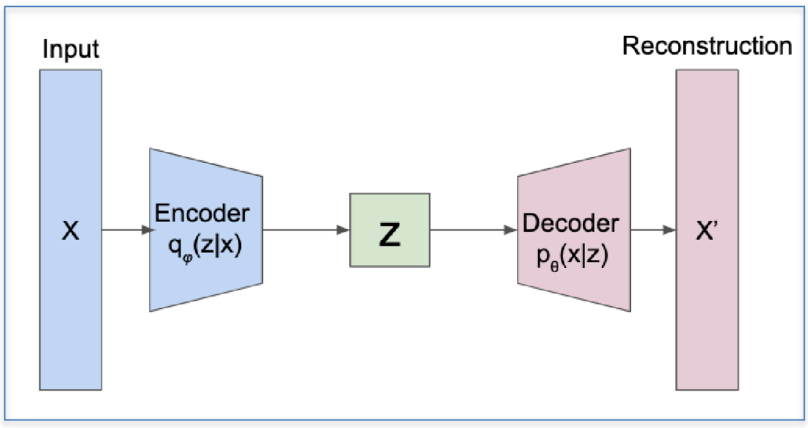

Figure 3. Basic Autoencoder architecture (Sikka, 2020)

utilizing a comprehensive attribute classifier, demonstrating the model's adaptability and efficiency across varied transformations. The another GAN-based method (Shao et al., 2021), which successfully conveys the emotional state in images with different human face poses and angles in the image, extracts facial features from two unaligned images with an encoder pair. A disentanglement and generation network represents the differences between these features as a vector. Afterwards, these values are swapped with each other with another encoder network. Finally, the output image is obtained with a generator network.

A GAN is presented in Fig. 2 consists of two components: a generator that creates synthetic data from random noise, attempting to mimic the distribution of real data, and a discriminator that evaluates whether the data is real (from the dataset) or fake (produced by the generator). Figure 3 shows an Autoencoder which is a neural network composed of an encoder, which reduces the input data to a lower-dimensional

latent space, and a decoder, which reconstructs the input data from this compressed representation, typically used for tasks like dimensionality reduction, denoising, or feature learning. Most of the deep learning-based methods described below are built on GAN or Autoencoder architectures.

2.2.1. StarGAN V2

The StarGAN v2: Diverse Image Synthesis for Multiple Domains (Choi et al., 2020) paper enhances the original StarGAN model (Choi et al., 2018), significantly improving the diversity and quality of image generation across various domains, with notable applications in facial attribute transfer. This advanced model consists of a Generator that produces images by transforming a source image and a style code, a Discriminator that differentiates between real and fake images, a Mapping Network that generates diverse style codes from latent codes and domain labels, and a Style Encoder that facilitates style transfer between images. StarGAN v2 excels in facial attribute manipulation by allowing users to specify target attributes and achieve direct style transfers from reference images, maintaining the identity of the original while altering specific features. Its training incorporates adversarial, style reconstruction, and cycle consistency losses to ensure realistic outputs. The model's capability to control attribute intensity and combinations makes it ideal for tasks like aging simulation, emotion synthesis, and identity preservation in facial editing, highlighting its utility in generating nuanced and controlled transformations in multiple domains.

2.2.2. STD-GAN

The STD-GAN (Guo et al., 2021) paper introduces a sophisticated method for instance-level facial attribute transfer, enhancing fidelity and style accuracy using binary attribute annotations. This model surpasses previous methods by efficiently capturing the nuances of style in attribute transfers. STD-GAN operates through a two-step process: first removing the original attribute from the target image, then adding the new attribute with its style extracted from a source image, facilitated by a Style Disentangling Module. The architecture features a generator for image creation

and a discriminator to evaluate the authenticity and style correctness, along with a style encoder for precise style capture. Training incorporates adversarial methods, attribute classification loss, and style regression loss to ensure the generated images are realistic and maintain the intended attributes and style. The model's ability to perform semantic-level attribute editing broadens its applicability in image editing, synthetic image generation, and personalized content creation, significantly advancing attribute transfer techniques.

2.2.3. Cascade EF-GAN

The Cascade EF-GAN: Progressive Facial Expression Editing with Local Focuses (Wu et al., 2020) paper introduces an innovative Generative Adversarial Network (Goodfellow et al., 2014) approach for facial expression editing, specifically designed to tackle the challenges of artifacts and blurring in expression-rich areas during large-gap transformations, such as changing from a furious to a laughing expression. The Cascade Expression Focal GAN (Cascade EF-GAN) uses a series of EF-GAN modules, each comprising an Expression Transformer for initial edits and a Refiner for final adjustments, focusing on local facial features (eyes, nose, mouth) to preserve identity and enhance realism. This progressive, multi-step approach reduces artifacts and maintains detail by treating global and local edits separately. Key innovations include a cascade strategy for gradual transformation, local focuses for detail preservation, and advanced loss functions like adversarial, expression, content, attention, and interpolation loss to refine output quality. This model represents a significant leap forward in GAN-based image editing, offering potential applications in broader image editing tasks while maintaining the integrity of the original images.

2.2.4. Latent Space Mapping for Complex Attribute Synthesis

Latent space mapping techniques (Nitzan et al., 2020) , which involve mapping and manipulating latent representations of facial images, facilitate complex transformations, including aging, style changes, and more. These methods leverage the power of variational autoencoders (Kingma and Welling, 2019) or similar architectures to

explore and modify deep features that control specific facial characteristics. However, as we have observed in our experimental studies, when an image different from the training set is input into the system, the face most similar to those in the distribution will be synthesized. This synthesized face will not connect to the input image regarding personal resemblance as shown in Fig. 27.

2.2.5. Disentangled Representation for Expression Transfer

Disentangled representations allow users to manipulate specific facial attributes independently. GeneGAN (Zhou et al., 2017) effectively separates attributes such as eyeglasses or smiles into distinct subspaces, enabling precise control over modifying chosen features while preserving the subject's identity. The GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data (Zhou et al., 2017) paper presents a novel use of generative adversarial networks (GANs) for object transfiguration in images, focusing particularly on facial attribute transfer. The model excels in tasks such as replacing eyeglasses, altering facial expressions, or changing hairstyles, utilizing unpaired datasets—one with the object and another without. GeneGAN's architecture features an Encoder and a Decoder; the Encoder separates an image into background and object features, while the Decoder can reconstruct the original image or craft new images by mixing features from different sources. This process is enhanced through adversarial training with a discriminator that ensures the realism and correct domain alignment of generated images. A standout feature of GeneGAN is its capability to handle unpaired data, obviating the need for directly corresponding 'with and without object' images and enabling the model to manipulate object presence robustly. The model also discerns and manipulates attribute subspaces complex representations of specific attributes like eyeglasses or smiles, allowing for precise and diverse image modifications. GeneGAN's potential extends to virtual try-on systems, automated photo editing, and more, representing a significant leap in image modification technology using unpaired data and GANs.

2.2.6. Advanced Techniques in Photorealistic Facial Texture Transfer

The study Photo-realistic Facial Texture Transfer (Kaur et al., 2017) tackles the complex task of transferring facial textures between images while preserving the original identity of the subject. The technique employs a combination of Markov Random Field (MRF) and Convolutional Neural Networks (CNNs), introducing a facial semantic regularization strategy that consists of facial prior regularization and facial structural loss. This method effectively preserves critical facial features such as the eyes, nose, and mouth during texture transfer, ensuring the identity remains intact. It outperforms traditional style transfer methods by maintaining both photorealism and semantic accuracy.

2.3. Innovations in Real-time Facial Reenactment

In the studies Thies et al. (2016) and Garrido et al. (2014) develop a groundbreaking approach to animate facial expressions in real time. The method leverages a deformation transfer function that efficiently operates within a low-dimensional expression space, facilitating instantaneous application. An integral part of their approach is using a 3D morphable face model that allows for precise tracking and re-rendering facial expressions to match real-world lighting and the ambient video setting seamlessly. This capability transforms live video feeds, enabling real-time applications such as video dubbing and enhanced teleconferencing, where facial expressions are altered dynamically without noticeable disruptions or loss of realism.

Integrating these diverse methodologies enhances the ability to perform sophisticated facial manipulations, from real-time expression changes to detailed texture and attribute editing. The ongoing developments indicate a strong movement towards models that adjust facial expressions and seamlessly integrate various facial attributes, ensuring photorealism and identity preservation.

The method presented in this article stands out among other approaches in the literature because it is a lightweight approach that is independent of the dataset, works fast and minimizes unwanted degradation. However, when non-deep learning-based approaches are considered, it is predicted that our method's ability to manipulate the

face while preserving facial identity will increase the preference rate of our method.

CHAPTER 3: METHODOLOGY

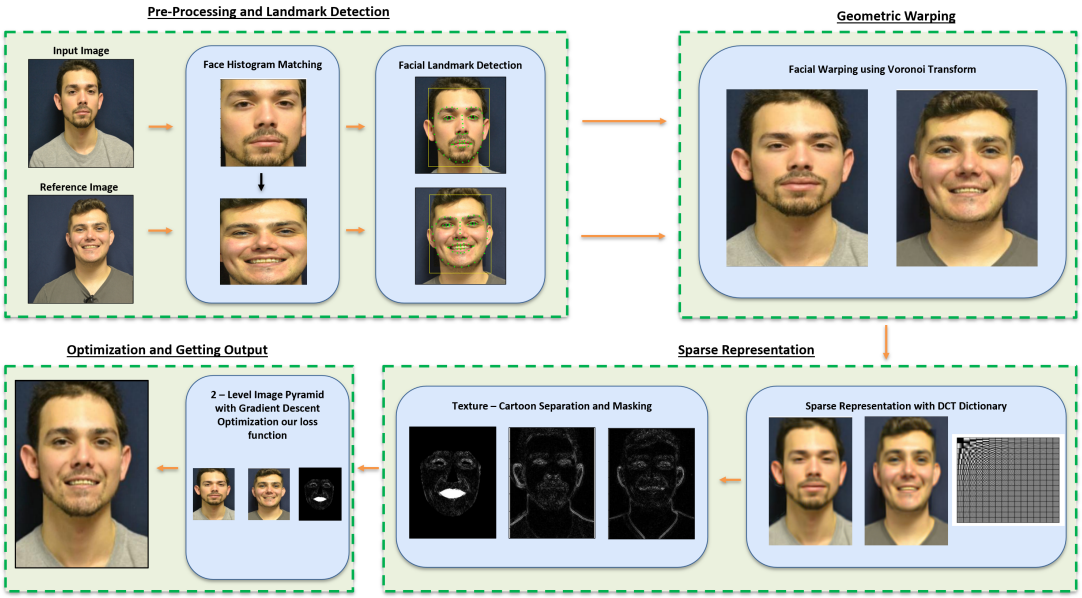

Figure 4. General overview of our method

Our method receives two input images and detects human faces within these images, followed by the identification of facial landmarks. Due to the diversity in human skin tones, color histogram equalization is performed between the two images, considering only the facial frames. Subsequently, the facial areas are prepared for warping through the Delaunay Triangulation and Voronoi Transform. Using facial landmarks, the overlay of human faces on each other not only aligns the images but also facilitates the transfer of emotional expressions (such as the fullness of the cheekbones or the enlargement of the eyes in fear). After overlaying, the images are converted into the HSI color space, and the intensity channel is divided into 8x8 overlapped patches. These patches are sparsely represented using a dictionary based on the Discrete Cosine Transform. The sparsely represented patches are reassembled to produce grayscale sparsely represented images. The difference between these sparsely represented and typical images is obtained using the mathematical subtraction operator. The resulting difference image represents the texture portion of Cartoon-Texture separation. This texture and the input and reference images are optimized at each level of a two-level image pyramid using the gradient-descent algorithm with the loss function provided in

our method. The relevant flowchart is presented in Fig. 4.

3.1. Color Histogram Matching

For the transfer we will make between human face images to be consistent, the pixel values between the input image and the reference image must be compatible. Differences in pixel values arising from people's skin colors, the environmental conditions under which the photos are taken, and the devices used to capture the images will cause inconsistencies in the pixel values of the image we obtain in the method's output. Therefore, the first step we implement in our method is to match the color histograms of the two images.

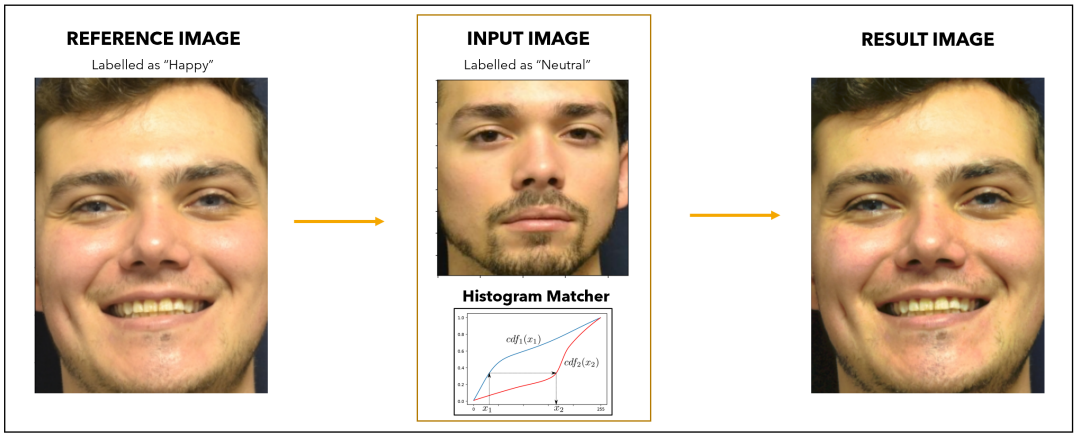

Figure 5. Reference image after applying color histogram matching

The histogram matching process as shown in Fig. 5 , applied only by considering the bounding box coordinates produced by face detection, has yielded consistent results by excluding the pixel values outside the face region of the reference image from the histogram editing process.

3.2. Face and Landmark Detection

The method we presented utilizes pre-trained models for the initial steps of our study, which include face detection and facial landmark detection. The RetinaFace (Deng et al., 2019) library was employed for face detection, while the DLIB (Kazemi and Sullivan, 2014) library was used for facial landmark detection. The model released

by DLIB, which detects 68 facial landmarks, was chosen.

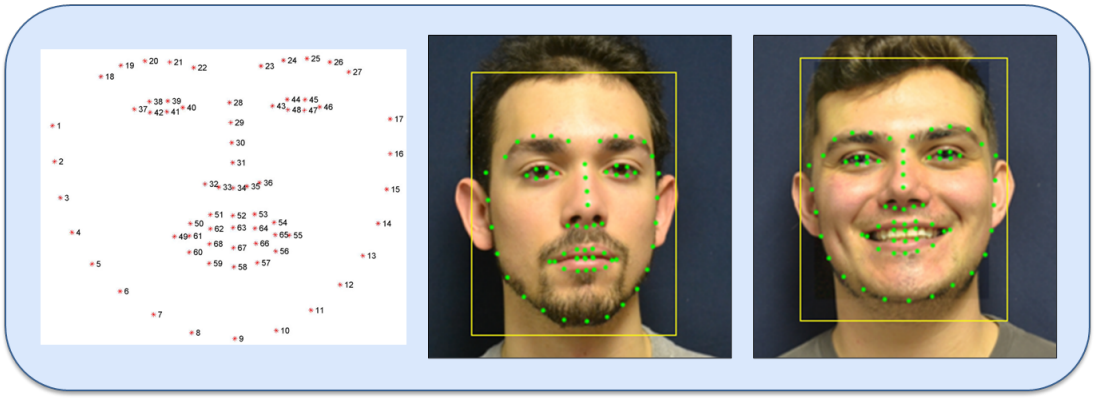

Figure 6. Landmark Points of DLIB Model (left), Facial Landmarks of Input Image (middle), Facial Landmarks of Reference Image (right)

In Fig. 6 , the landmark points corresponding to the landmark ID numbers of the model shown on the far left are displayed on the input image and the reference image, respectively.

3.3. Facial Warping Transform

A face-warping technique is applied to ensure that the optimization process is carried out as intended and to convey human emotions. To implement the warping geometrically, it is necessary to divide the facial areas according to a specific mathematical model. Therefore, the Voronoi Diagram, a derivative of the Delaunay Triangulation method (Lawson, 1972), is preferred. As shown in 7 , the facial region boundaries are divided into areas that are equidistant to neighboring points.

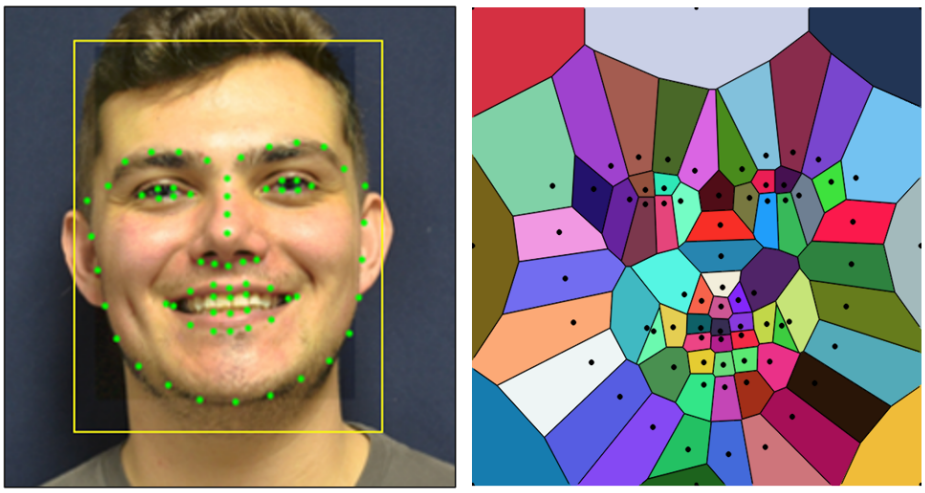

Figure 7. Voronoi Diagram Spaces according to input facial landmarks

The regions shown with unique colors on the right in Fig. 7 represent the regions containing separating edges equidistant from neighboring points, that is, the output of the Voronoi transform. Dividing the human facial region into these regions is a crucial step that allows us to apply the warping transform. This transform plays a significant role in transferring features such as puffy cheeks due to emotional state between images. Moreover, it ensures that the optimization phase works correctly by aligning the images on top of each other.

Through the warping process, while the facial regions of the input and reference images are aligned on top of each other, features such as the protrusion of the cheeks, the widening of the eye sockets, and the downward movement of the chin are geometrically transferred between the two images. As a result, the images were prepared for the optimization process and geometrically transformed to more consistently reflect the emotional state visually. The outputs of the warping transform are shown in Fig. 8

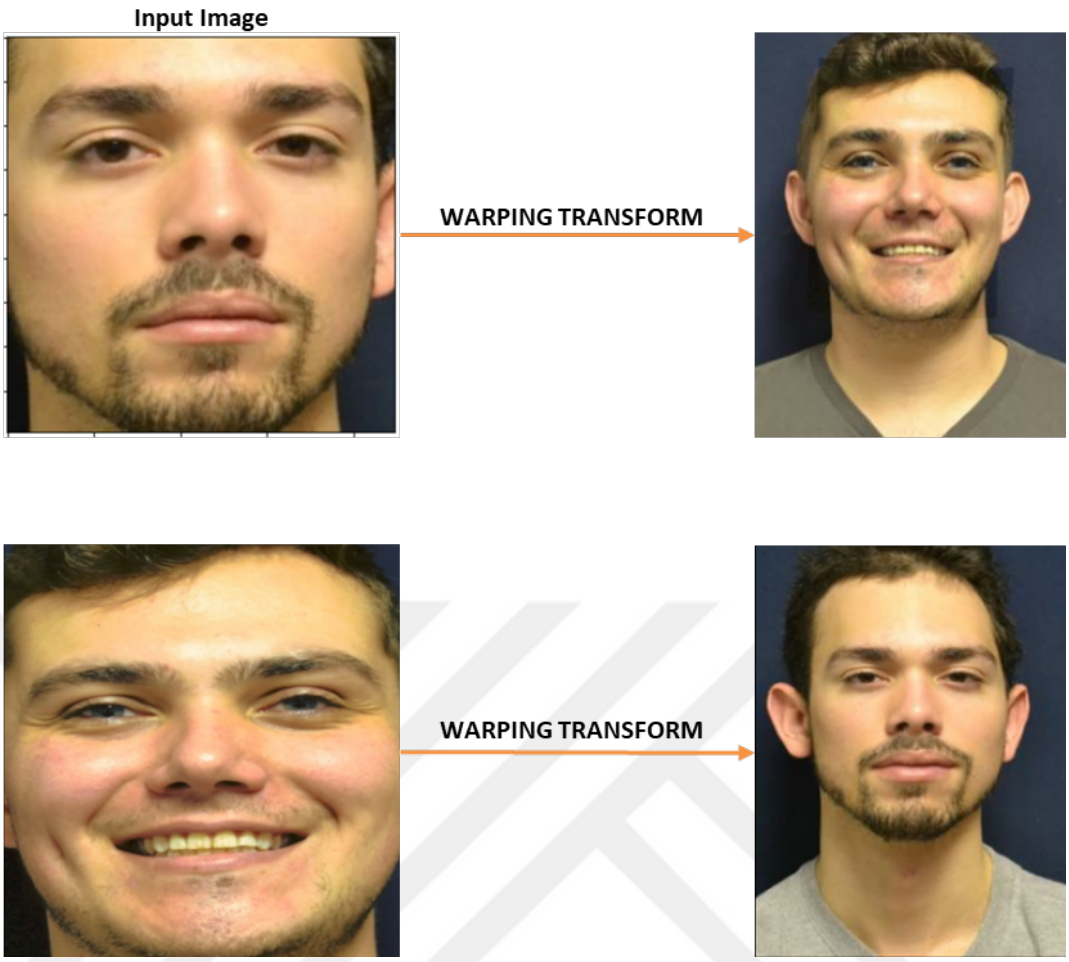

Figure 8. Input and reference images after applying the warping transform

3.4. Sparse Representation

Before optimally fusing two images, it is essential to determine the features we want in the output image. Specifically, subtle details like wrinkles around the eyes, other wrinkles, and dimples on the face can alter human emotional states and age-related characteristics. Artificial intelligence-based approaches can be used to extract these features from images. However, the process of locally labeling this data, due to its complexity, can pose a significant challenge. For these reasons, in the method presented in this article, sparse representation is used for cartoon-texture decomposition (Elad et al., 2005), and the results obtained are considered as the desired features. The steps for this process are outlined below in the subheadings.

3.4.1. *Initial Data Representation*

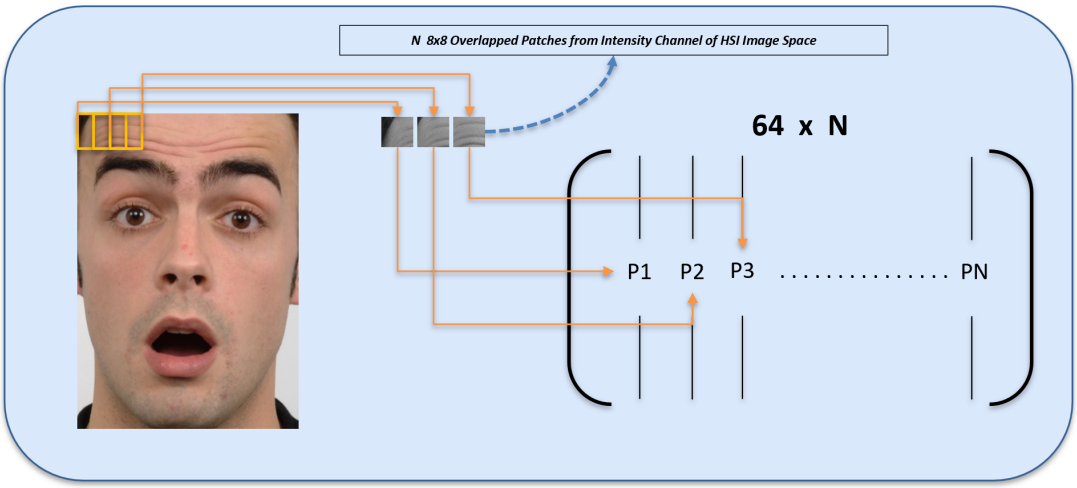

Figure 9. Representing the image in a matrix form

Before the input images were represented sparsely, these images were converted into a specific matrix format. While preparing this matrix, the input images were converted from the RGB color space to the HSI color space. The HSI color space's Intensity (I) channel was taken and divided into 8x8 overlapped N patches. Subsequently, these patches were represented in a $\mathbb{R}^{64}$ vector space. The represented vectors were concatenated in columns to form a matrix with dimensions of 64xN.

Figure 9 describes how, after converting the detected face region into HSI image space, the Intensity channel is divided into overlapped patches, each patch is converted into a vector, and the image is presented in a matrix form.

3.4.2. *Discrete Cosine Transform Dictionary*

Since we wanted to solve the problem with a general approach independent of the dataset, we opted to use a Discrete Cosine Transform-based dictionary that could represent the entire space of natural images. The dictionary we created consists of 256 patches, each of size 8x8.

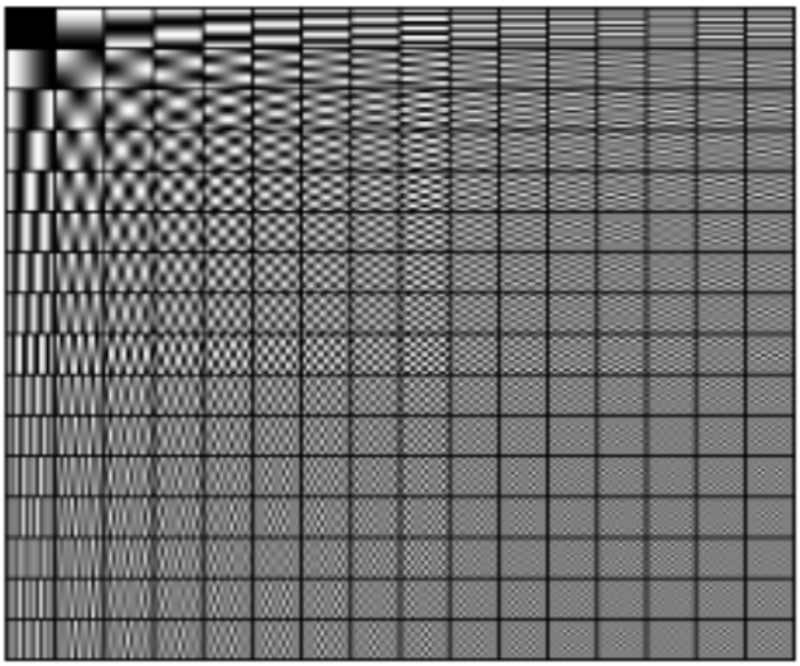

Figure 10. DCT Dictionary with 256 Patches

Figure 10 illustrates basis patches included in a DCT-based fixed dictionary. The dictionary contains 256 patches and is the essential dictionary we use to determine which basis the patches consist of with OMP (Cai and Wang, 2011), which will be explained in the next step.

3.4.3. Orthogonal Matching Pursuit (OMP)

In the equation $Dx = b$, which presents infinitely many solutions for x, the Orthogonal Matching Pursuit (OMP) (Cai and Wang, 2011) method was strategically chosen. This method, known for offering a least-squares solution within a subset, plays a crucial role in indicating how the given patch can be best represented by elements of the DCT-based dictionary.

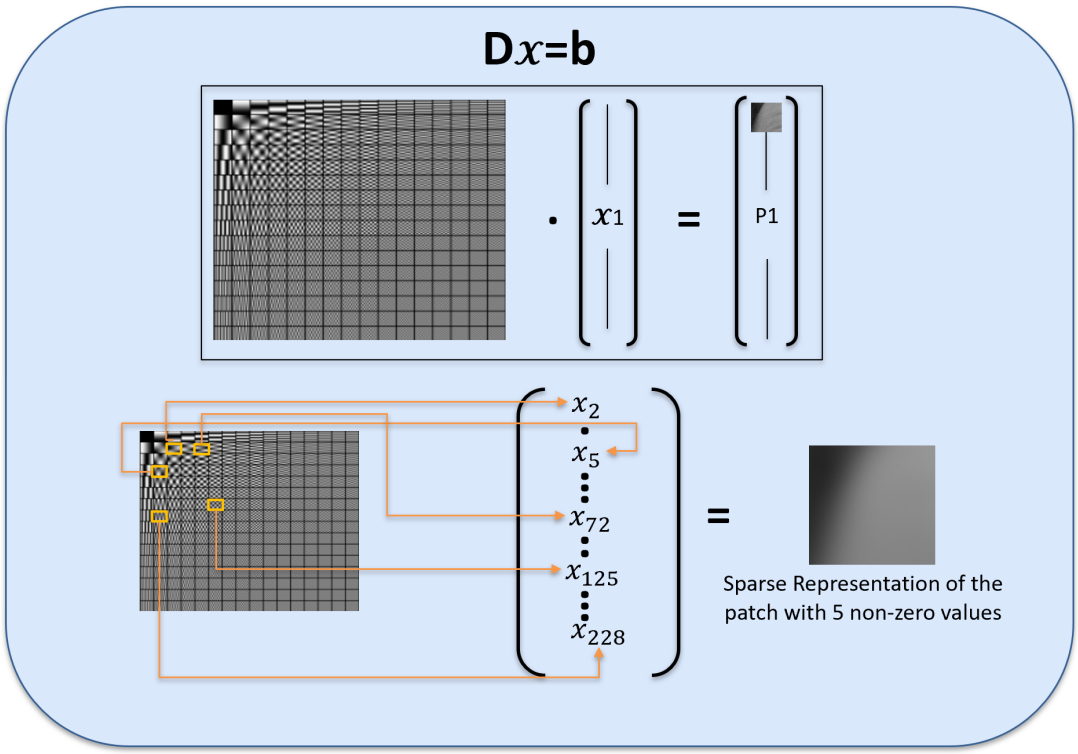

Figure 11. The Sparse representation process off all patches with sparsity is equal to 5, using DCT dictionary

Experimentally, we found that having five non-zero values in the solution vector (sparsity=5) provides the most optimal representation of human facial texture. After all overlapped patches are obtained sparsely, they are combined to achieve a sparse representation of the image. This process of preparing the image for the next step, the cartoon texture separation, is a crucial part of our methodology.

Figure 11 shows how a single patch from the intensity channel is represented as sparse using a DCT-based dictionary. Five non-zero values are calculated with OMP, and the relevant patch is represented as sparse. Then, the sparse is stored again as a vector. This process is applied to all extracted patches.

3.5. Texture - Cartoon Separation and Masking

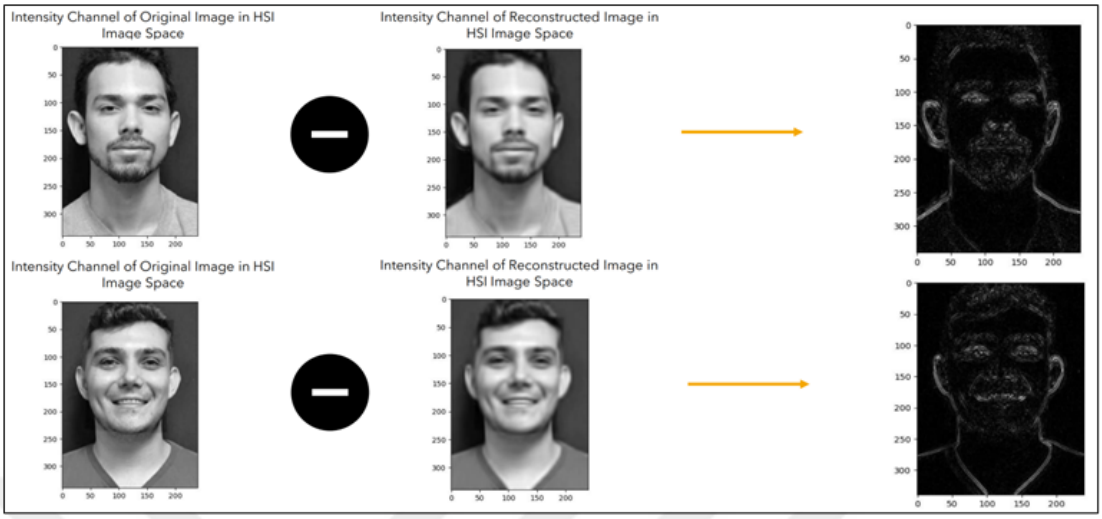

Figure 12. Applying the subtraction operator to get the texture part

When the sparsely represented versions of the input and reference images are subtracted pixel by pixel using a mathematical subtraction operator from their complete versions, the texture parts of the face images are obtained. A mask of the desired pixels is created by filtering out the pits, wrinkles, and lines at the corners of the eyes caused by smiling on the faces. Moreover, since the goal is to convey human emotions by transferring the entire mouth area, the mouth region is filled with 255 (white) pixel values according to the facial landmark coordinates.

Figure 13. Texture part of the reference image (left), the filtered redundant parts and areas of the texture with a white-masked mouth, black-masked eyes, and black-masked nose (right)

In Fig. 12, texture information is obtained by subtracting the intensity channel of the input and reference images and the sparse versions of these channels. The texture information obtained in Fig. 13 was filtered according to facial landmark coordinates and removed from unwanted areas. This way, the optimization process will be initiated by considering the person's mouth structure, cheeks, wrinkles, and scars around the eyes.

3.6. Optimization

The developed method utilized stochastic gradient descent optimization to obtain the final image. The designed combined loss function was minimized using gradient descent to produce the output image. Mathematically, the Combined Loss consists of the linear combination of three different loss functions. The coefficients of this linear function were determined within the scope of experimental studies. Moreover, each step of the optimization process was applied to every layer of a two-layer image pyramid system created for the input, reference, and mask images, ensuring

consistency in small details and rapid optimization in the images.

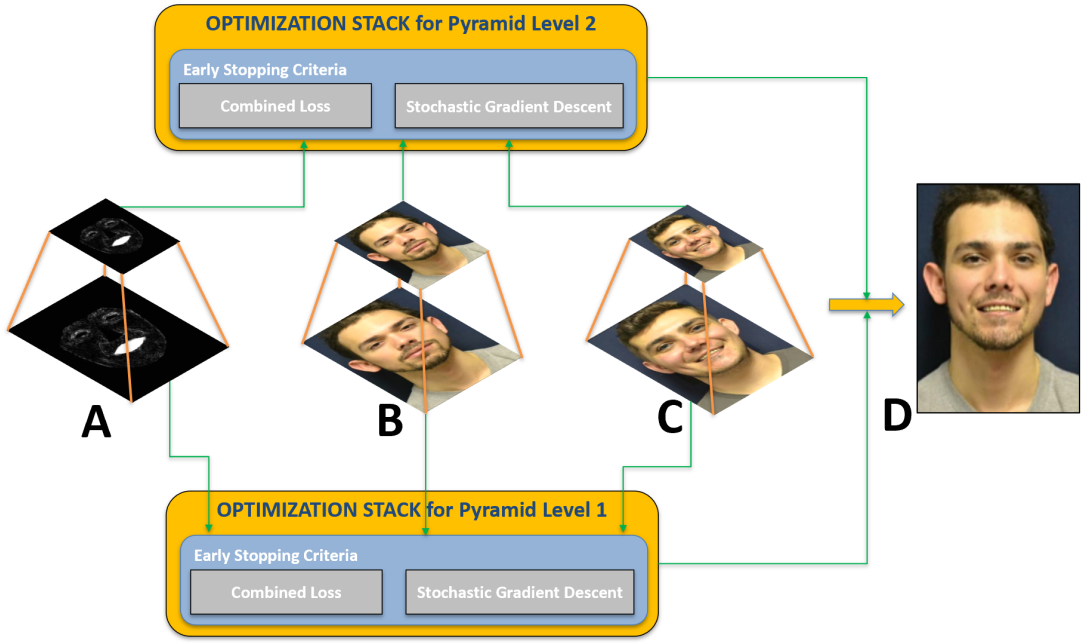

Figure 14. The Optimization flow using image pyramids, **A**: Mask Image, **B**: Geometrically Warped Input Image, **C**: Geometrically Warped Reference Image, **D**: Optimized Output Image

Figure 14 shows how the optimization process is optimized using input, reference, and filtered mask images. A two-level image pyramid structure ensures effective optimization of small details and quick optimization. The images on the second level and the images on the first level pyramid are included in the optimization process according to their own scale values.

The Combined Loss consists of the linear combination of the Mean Squared Error Loss (MSE), Total Variation Loss (TV) (Rudin et al., 1992), and Color Distribution Loss (CD) (Eq. 3) functions. The Mean Squared Error Loss function penalizes high differences between two images. Thanks to the MSE component of the Combined Loss, the difference obtained by subtracting the geometrically distorted input and reference images is multiplied by the mask image obtained from the sparse representation process, ensuring that the feature regions found in the reference image are placed within the input image. In the equations Eq. 1, Eq. 2, and Eq. 3, **A** represents the input image that has been geometrically warped, **B** represents the

reference image that has also been geometrically warped, and **M** represents the mask image obtained from the cartoon-texture separation module. Additionally, the terms H and W in these equations represent the height and width of the images (which are the same for all images) provided as input to the function. Moreover, the term b in equation Eq. 3 is used to denote the pixel values in the histogram (all pixels are processed in the presented method).

$$MSE(A,B,M) = \frac{1}{\sum M} \sum_{h=1}^{H} \sum_{w=1}^{W} M_{hw} \cdot (A_{hw} - B_{hw})^2 \tag{1}$$

Among the other components of the Combined Loss, the Total Variation Loss in equation Eq. 2 minimizes signal noise in the image, while the Color Distribution Loss in equation Eq. 3 helps maintain color consistency by minimizing the absolute difference between the histograms of the input and reference images. The image used in the Total Variation loss in equation Eq. 2 function is the constantly updated input image.

$$TV(A) = \sum_{h=1}^{H-1} \sum_{w=1}^{W-1} \sqrt{(A_{h+1,w} - A_{h,w})^2 + (A_{h,w+1} - A_{h,w})^2} \tag{2}$$

$$CD(A,B,b) = \sum_{i=1}^{b} |\text{Hist}_A(i) - \text{Hist}_B(i)| \tag{3}$$

The Combined Loss shown in Equation 4 and 5, which consists of a linear combination of three different loss functions, has been minimized using gradient descent optimization with experimentally determined α, β, and γ parameters to obtain the most optimal image. The number of iterations for optimization has been dynamically adjusted such that the iteration completes if the difference in loss values between two consecutive iterations is not greater than a particular epsilon value or if the loss value increases, using the early stopping method.

$$\text{minimize} \quad Loss_{\text{combined}}(A,B,M) \tag{4}$$

$$Loss_{\text{combined}}(A,B,M) = \alpha MSE(A,B,M) + \beta TV(A) + \gamma CD(A,B,b) \tag{5}$$

Algorithm 1: Multi-Scale Optimization with Gradient Descent

input : Image A, Image B, Mask
output: Optimized Image **current_image**
begin
  Build Pyramid $\mathbf{P_A}$ for **Image A (input image)**;
  Build Pyramid $\mathbf{P_B}$ for **Image B (reference image)**;
  Build Pyramid $\mathbf{P}_{\text{mask}}$ for **Mask (mask image)**;
  `// Initialize with the coarsest level of Image A`
  current_image $\leftarrow \mathbf{P_A}[-1]$;
  for $level = 1$ *to* L **do**
    for $iteration = 1$ *to* N **do**
      current_image resized to match size of $\mathbf{P_B}[level]$;
      $\nabla \leftarrow$
       gradient of combined_loss(**current_image**, $\mathbf{P_B}[level], \mathbf{P}_{\text{mask}}[level]$);
      $\nabla \leftarrow$ L1 regularization$(\nabla, \lambda_{\text{reg}})$;
      current_image $\leftarrow$ **current_image** $- \eta \times \nabla$;
      current_loss $\leftarrow$ Calculate combined_loss;
      if *Early Stopping Condition Met* **then**
        break;
      end
    end
    if $level < L$ **then**
      current_image $\leftarrow$ Upscale to next level;
    end
  end
  return current_image;
end

In order to optimize the image using the designed loss function, as shown in Algorithm 1, L1 regularization (Tibshirani, 1996) is applied after taking the gradient of these loss functions. L1 regularization is applied to take a robust approach against noise, accelerate optimization, and obtain a realistic image at the output by increasing the method's generalization ability. In Algorithm 1 η means learning rate, L means total level count in the image pyramid and, ∇ means gradient. Also, λ_{reg} parameter is essential for regulating the impact of L1 regularization in the method's complexity and its ability to generalize.

CHAPTER 4: EXPERIMENTAL RESULTS

The experimental studies of the method were conducted using the TUFTs RGB Emotion (Panetta et al., 2020) and RafD (Langner et al., 2010) datasets. The TUFTs RGB Emotion dataset contains front-facing images of 113 individuals, including three emotional states (neutral, happy, and surprised). The RafD dataset provides eight different emotional states for 78 different individuals. In the experimental studies, the performance of the emotion classifier was measured based only on the neutral, happy, and surprise emotional states. Four hundred forty-eight simulations were carried out, with 336 using the TUFTs RGB Emotion dataset and 112 using the RafD dataset.

4.1. Metrics

The Fréchet Inception Distance (FID) (Heusel et al., 2017) is a crucial metric for assessing the quality of images produced by models like GAN (Goodfellow et al., 2014), quantifying how well generated images resemble real images in terms of their distribution. FID involves passing both real and generated images through the Inception v3 (Szegedy et al., 2015) network to extract feature vectors, which are then used to model each image set as a multivariate Gaussian distribution with mean μ and covariance σ. The FID score is calculated using the formula which is shown in Eq. 1. FID Score, where lower scores indicate more similarity and higher image quality. This metric effectively captures both the diversity and fidelity of the generated images relative to the actual ones.

$$FID = \|\mu_{real} - \mu_{gen}\|^2 + \mathrm{Tr}(\Sigma_{real} + \Sigma_{gen} - 2(\Sigma_{real}\Sigma_{gen})^{1/2}) \tag{1}$$

The term $\|\mu_{real} - \mu_{gen}\|^2$ in Eq. 1 represents the squared Euclidean distance between the mean vectors of the real μ_{real} and generated μ_{gen} image distributions. The mean vectors are calculated from the feature vectors extracted by the Inception v3 (Szegedy et al., 2015) model. This distance measures how far apart the central tendencies of the two distributions are, with a smaller distance indicating more

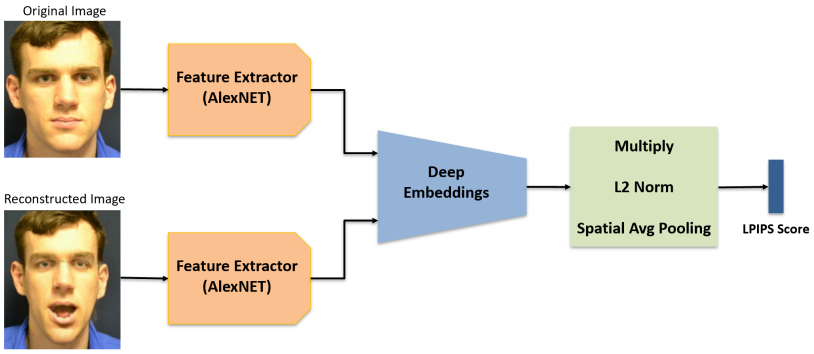

Figure 15. General structure of LPIPS

similarity in the central features of the images.

These terms, Σ_{real} and Σ_{real}, in Eq. 1, are the covariance matrices of the feature vectors for the real and generated images. Covariance matrices capture the spread and the correlation structure of the features within each set of images.

Term $\Sigma_{real} + \Sigma_{gen}$ in Eq. 1 is the sum of the covariance matrices of the real and generated images. It represents the total variability within each set of images. Term $(\Sigma_{real}\Sigma_{gen})^{1/2}$ in Eq. 1 is the matrix square root of the product of the two covariance matrices. It serves to assess how similar the dispersion and orientation of the data points are between the two distributions.

The Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) score is a metric used to evaluate the perceptual similarity between two images (lower LPIPS score is better), such as a generated image and a ground truth image. This process involves passing both images through a convolutional neural network like AlexNet (Krizhevsky et al., 2012) to extract deep feature maps. The features of these two images are then normalized and subtracted to pinpoint differences. These differences undergo further L2 normalization and spatial averaging, enhancing significant discrepancies through multiplication by their L2 norms. The final LPIPS score aggregates these values, where a lower score indicates a closer perceptual resemblance to the ground truth.

Figure 15 shows LPIPS structure basically. Importantly, LPIPS aligns more with human visual judgment than traditional metrics like Peak Signal-to-Noise Ratio (PSNR) (Horé and Ziou, 2010), demonstrating its superiority. This makes LPIPS particularly valuable in fields like image generation and texture transfer where visual

fidelity is crucial.

Another metric used to measure the performance of the method is the score of the emotion classifier. In addition to obtaining a visually manipulated image with the method, determining the emotional state of the resulting image is also one of the method's objectives. To ensure fairness in generating the classification score, a globally published emotion classification model (Ryumina et al., 2022) has been used, and the relevant results are shared in this article. However, the other methods we compared have trained a classifier model using both ground truth images and generated images from their datasets for their classification scores. Therefore, for the approach mentioned in this article, a pre-trained ResNet18 model (He et al., 2015) has been fine-tuned. This was done by training the model on ground truth images from both the RafD (Langner et al., 2010) and TUFTs Emotion (Panetta et al., 2020) datasets, as well as generated images obtained using both datasets. The results of the trained classifier model are presented in Table 2, and the results obtained with an independent publicly available emotion classifier model are presented in Table 1 and Table 3. The independent emotion classifier model used (Ryumina et al., 2022) is a model trained on the AffectNET (Mollahosseini et al., 2019) dataset with additional data . Also, another metric we use is whether the probability percentage of the target emotional state, which the reference image possesses and we aim to achieve in the input image, has been increased in the classifier model compared to the initial state of the input image.

4.2. Results

Figure 16 and 17 underscore the importance of the neutral emotional state in our method by presenting the results of transforming sample pairs from different datasets from a neutral emotional state to a happy and surprised emotional state. These two scenarios are where the method works most successfully. The significance of the neutral emotional state is that the human face in the input image is as free as possible from wrinkles and facial dimples. This understanding is crucial as it allows for a more successful transfer of a higher-frequency image signal to a lower-frequency surface with our proposed method. Unlike the first two simulation scenarios, in the remaining

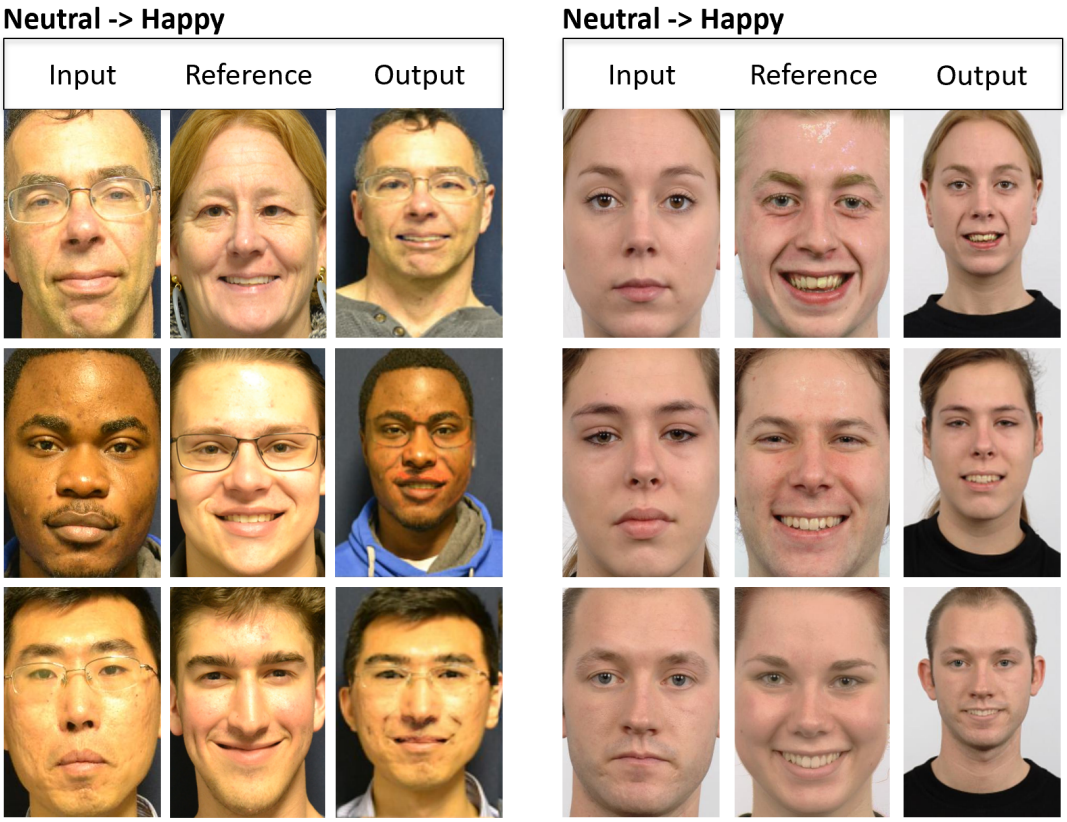

Figure 16. Generating the Output image from input images with a neutral emotional state using reference images with a happy emotional state, TUFTs Emotion (left), RafD (right)

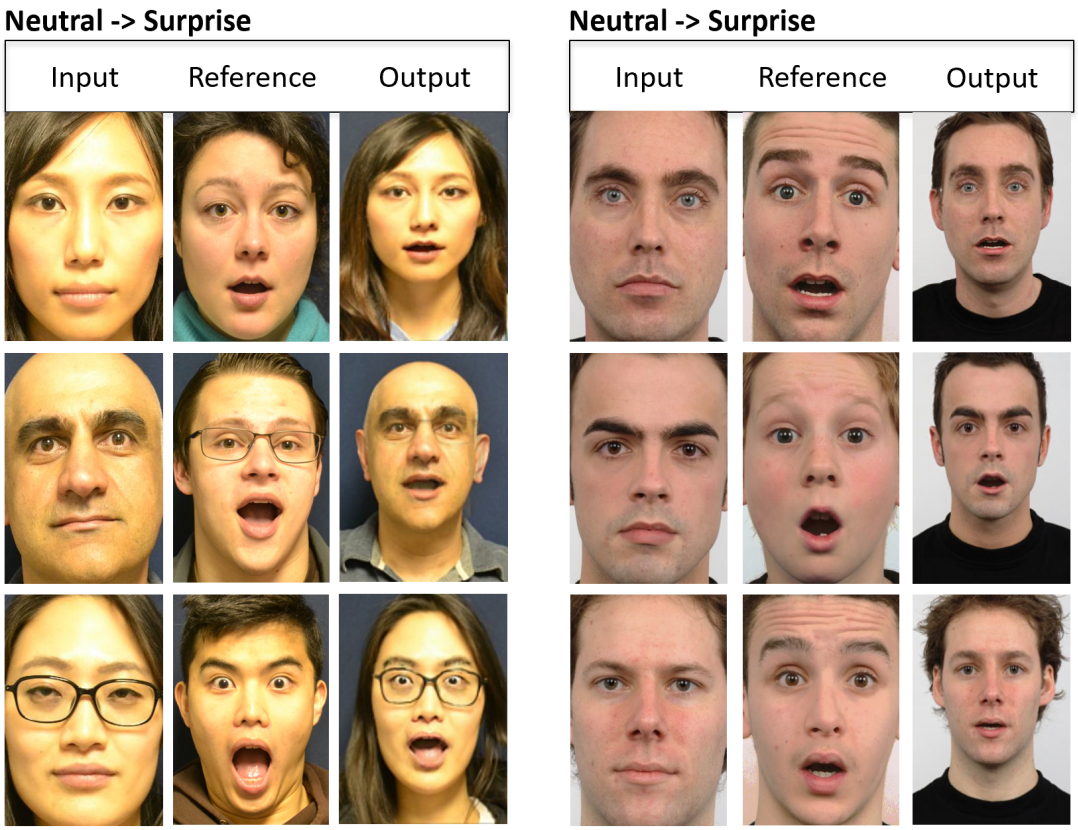

Figure 17. Generating the Output image from input images with a neutral emotional state using reference images with a surprise emotional state, TUFTs Emotion (left), RafD (right)

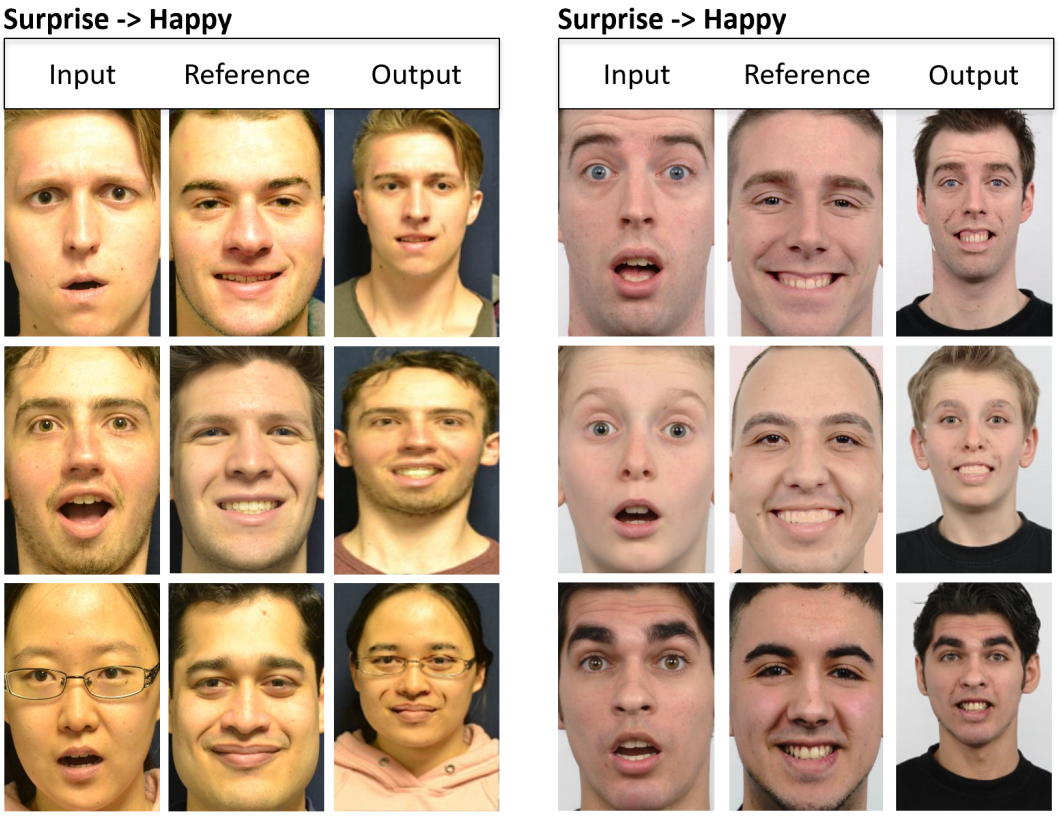

Figure 18. Generating the Output image from input images with a surprise emotional state using reference images with a happy emotional state, TUFTs Emotion (left), RafD (right)

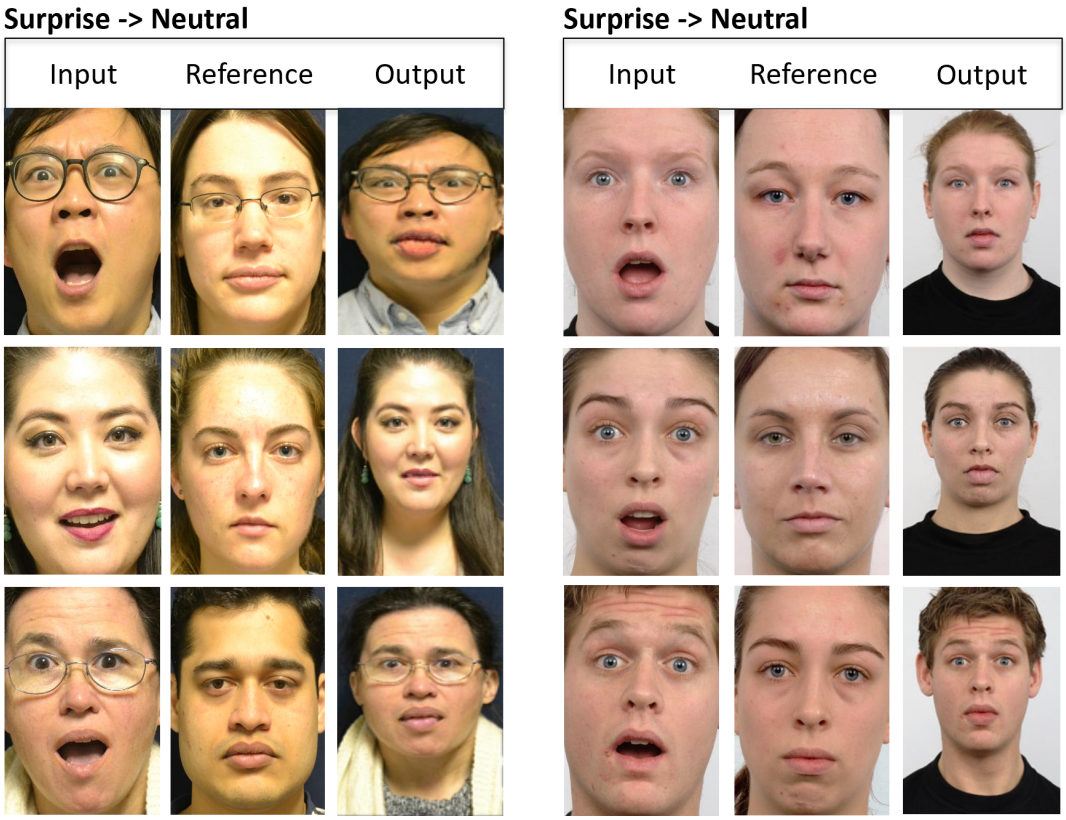

Figure 19. Generating the Output image from input images with a surprise emotional state using reference images with a neutral emotional state, TUFTs Emotion (left), RafD (right)

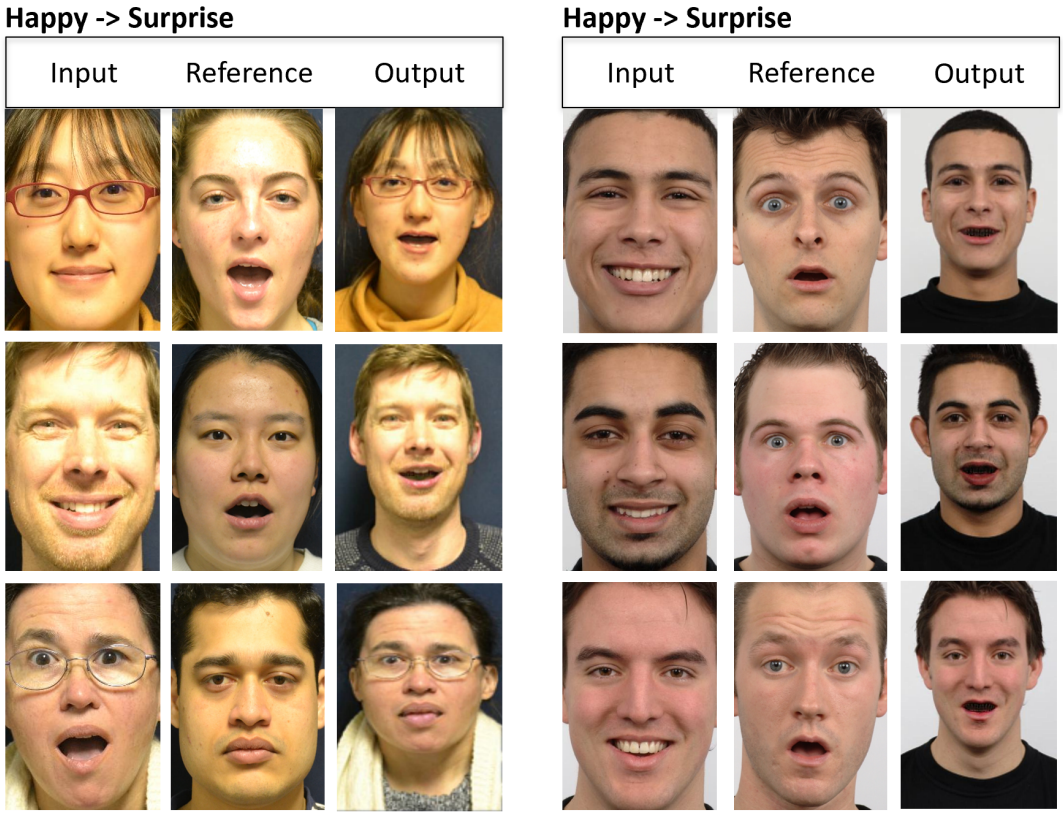

Figure 20. Generating the Output image from input images with a happy emotional state using reference images with a surprise emotional state, TUFTs Emotion (left), RafD (right)

scenarios, the proposed method is successfully applied to the images. However, the output images do not fully reflect the target emotional state, presenting a significant challenge in our research. This is an expected situation and has been included in the article to analyze and present the performance of the method in the best way. Our method does not promise the transfer of a texture with low frequency to a surface with high frequency. When a person smiles, wrinkles and dimples can form around the cheekbones and eyes. In Figures 19 and 21, when our method is run with the goal of converting from happy and surprised emotional states back to a neutral state, the results lead to the creation of human face images that appear more like someone struggling not to smile or displaying an anxious look rather than a truly neutral emotional state. Indeed, the low scores our method receives from a publicly available emotion classifier is an outcome of including these scenarios in the results.

However, as indicated in Fig. 18 and Fig. 20, while some test results are visually

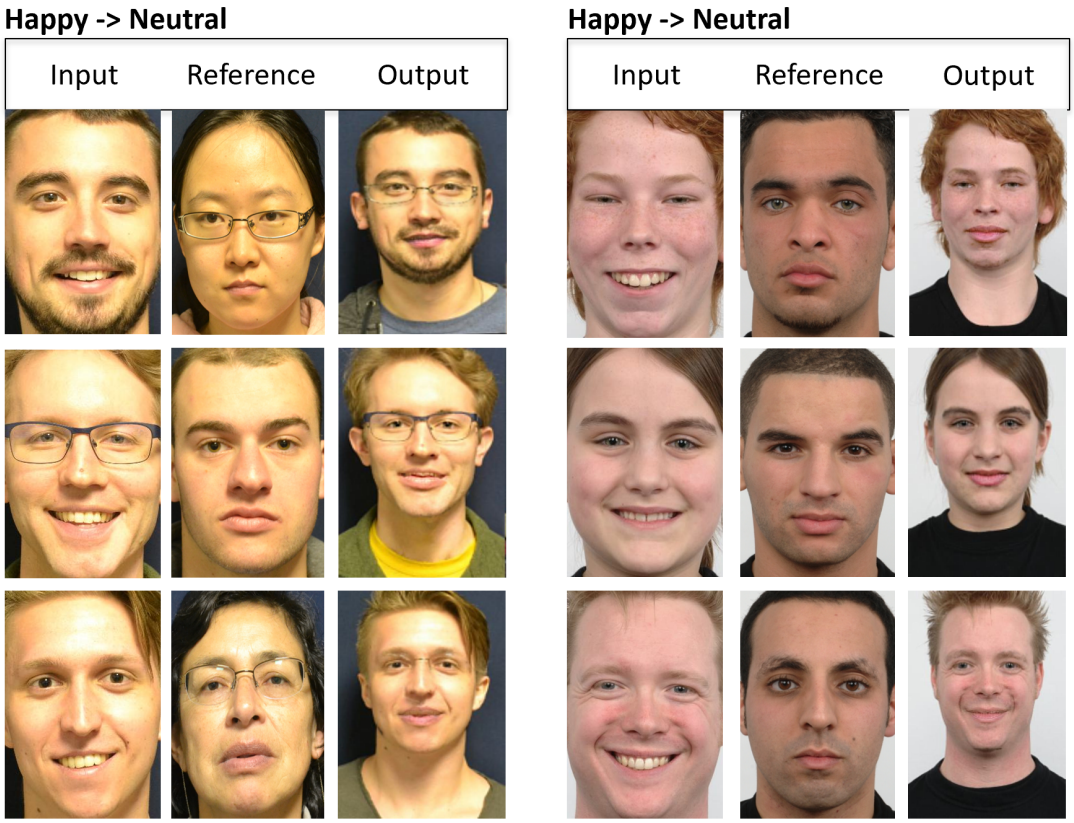

Figure 21. Generating the Output image from input images with a happy emotional state using reference images with a neutral emotional state, TUFTs Emotion (left), RafD (right)

	Datasets	
Metrics	RafD	TUFTs
PSNR	30.68	16.94
FID	6.90	15.45
LPIPS	0.02	0.21
Incrementation	0.91	0.87
Classification Score	0.30	0.41

Table 1. Overall Score table of proposed method

satisfactory when transitioning from a surprised to a happy emotional state, or vice versa, in extreme emotional states (such as extreme surprise), our method tends to produce outputs where the eyes are smiling but at the same time look surprised, or the facial expressions seem to be of someone trying to smile despite being surprised.

In Table 1, the overall results of the method, pairs of input and output images are used to show the Peak Signal-to-Noise Ratio (PSNR) (Horé and Ziou, 2010), Fréchet Inception Distance (FID) (Heusel et al., 2017), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), Incrementation, and Classification scores for two different datasets. The Classification Score is the success rate of transferring the emotional state from the reference image to the input image. It means that the emotional state in the reference image is also obtained in the output image. A publicly available Dynamic Facial Expression Recognition model (Ryumina et al., 2022) was used to calculate the Classification Score. The increment score indicates whether the percentage of the emotional state in the reference image is increased in the version of the input image before manipulation, and it is also calculated using the same model (Ryumina et al., 2022). Even if the emotional state in the reference image does not become dominant in the output image, a positive increment in its percentage positively affects the incrementation score. The proposed method increased the probability of the reference emotional state by an average of **0.35** in **0.91** of the simulations created using the RafD dataset and **0.87** in the TUFTs Emotion dataset. The outputs of the proposed method were compared with successful GAN models such as StarGAN (Choi et al., 2018), GANimation (Pumarola et al., 2020), and Cascade EF-GAN (Wu et al., 2020), which synthesize face images probabilistically and manipulate face images to create different emotional states in Fig. 22. This comparison was made

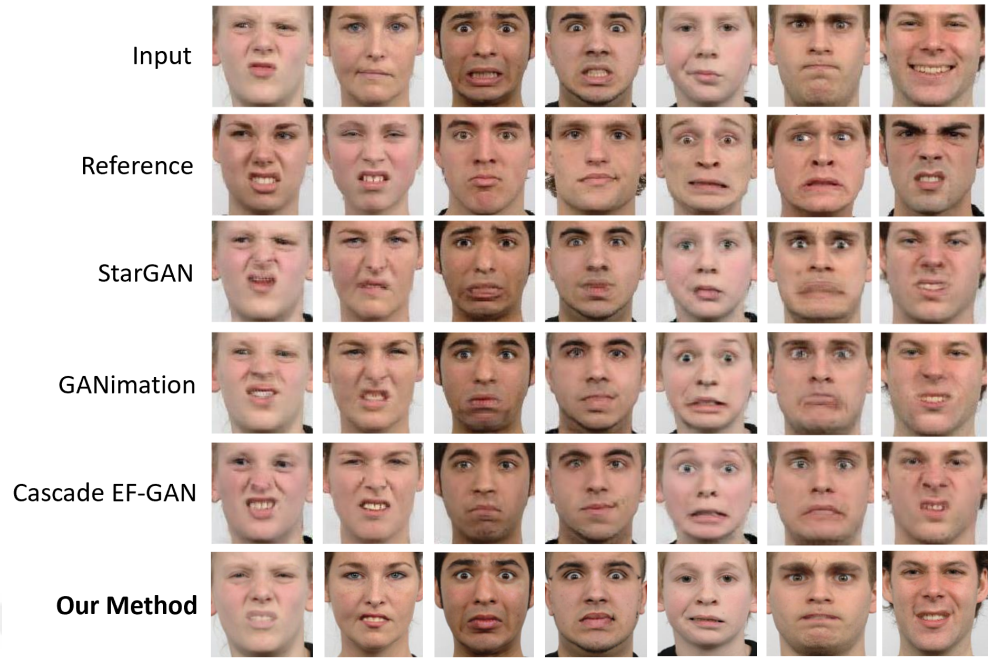

Figure 22. Visual Comparing Table with several GAN methods (Choi et al., 2018; Pumarola et al., 2020; Wu et al., 2020) using only RafD Dataset (Langner et al., 2010)

using the RafD dataset. The visual outputs other than the outputs produced by our method have been directly taken from the Cascade EF-GAN publication (Wu et al., 2020), due to the proprietary nature of some articles' codes. The same examples have been used to provide a fair comparison of our method's outputs, as shown in Fig. 22. Upon examining the visual results, it was observed that although the proposed method was ineffective in the upper parts of the eye region (forehead and eyebrows), it performed emotional state and facial texture transfer with minimal noise, producing results comparable to advanced GAN methods. The visual results table shows that GAN models sometimes caused inconsistencies in certain parts of the human face.

Method	**PSNR ↑**	**FID ↓**	**Classification Result**
StarGAN	19.82	42.39	0.884
GANimation	22.06	29.07	0.923
Cascade EF-GAN	23.07	27.15	**0.936**
Ours	**30.68**	**6.90**	0.807

Table 2. Score table of comparison between several GAN methods (Choi et al., 2018; Pumarola et al., 2020; Wu et al., 2020) and our method by using only RafD Dataset

In table 2 Our method stands out with a higher PSNR score (higher is better) and a lower FID score (lower is better) compared to other methods. Additionally, although the method does not excel in the success rate of emotion transfer, it still achieves a commendable percentage. The values in Table 2, titled 'Classification Results', have been directly taken from the Cascade EF-GAN publication (Wu et al., 2020). As indicated in the same publication, these results were obtained using an emotion classifier trained with both original and generated images from the Cascade EF-GAN publication. For the sake of a fair comparison, the authors of this paper have also trained an emotion classifier model using a dataset that includes both generated and original images with a similar approach, and obtained the classification results in this manner. Additionally, Table 1 presents results obtained independently using a publicly available emotion classifier (Ryumina et al., 2022), demonstrating an unbiased approach.

The visual results in Fig. 23 and metrics of simulations in Table 3 conducted using the TUFTs Emotion dataset for the scenario of transforming the emotional state from neutral to happy are provided for the comparison between the proposed method and the STD-GAN method (Guo et al., 2021), which produces satisfactory results in facial attribute transfer. The STD-GAN method facilitates the transfer of attributes such as bangs and eyeglasses using the respective keywords while transforming the person's image from a neutral to a happy emotional state using the keyword 'smile.' Since the common intersection of emotional state transformations for these two methods is only the transformation to a happy emotional state, simulations were run exclusively

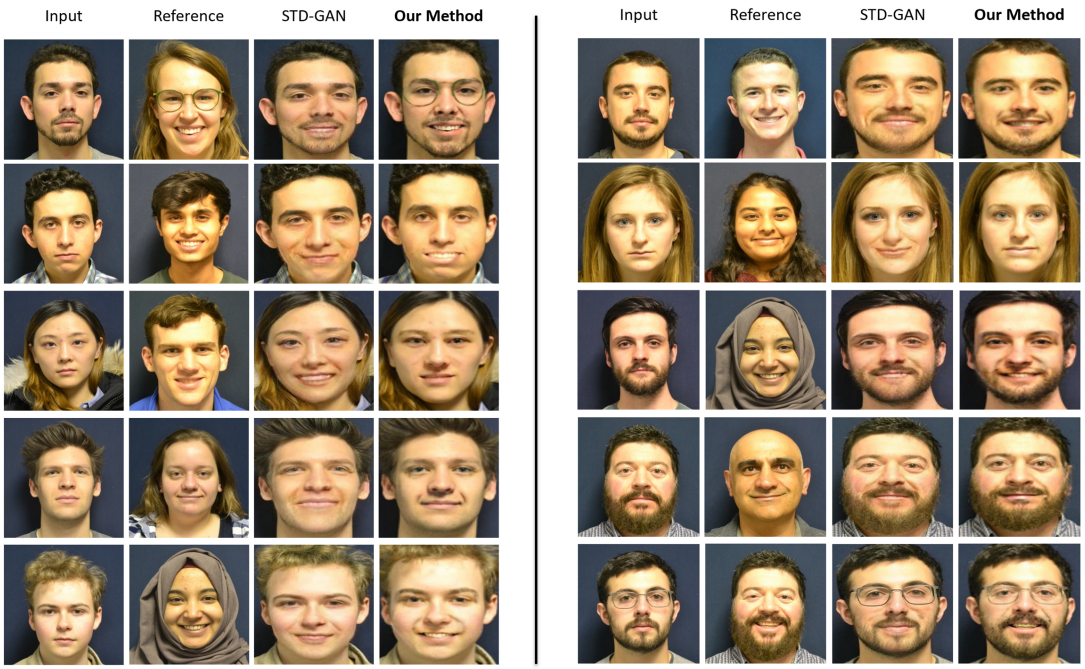

Figure 23. Neutral to Happy Simulation Results and Comparing with STD-GAN method, using only TUFTs Emotion Dataset

Method	PSNR	LPIPS	FID	Classification Result
STD-GAN	**18.42**	**0.17**	**11.55**	**0.94**
Ours	16.94	0.21	15.45	0.87

Table 3. Score table of comparison between STD-GAN (Guo et al., 2021) and our method by using only TUFTs Emotion Dataset and Neutral to Happy transformation scenario

for this scenario. Although the STD-GAN method leads in all metrics, the proposed method produced comparably successful results across all scores compared to the STD-GAN outputs. The Dynamic Facial Expression Recognition model (Ryumina et al., 2022) was used to obtain the classification results. Table 3 illustrates that our method, while trailing in all scores compared to the robust STD-GAN (Guo et al., 2021) , still manages to produce comparable results with GAN-based methods. This is due to its unique features of not requiring a dataset, being dataset-independent, and lighter in implementation. The 'Classification Result' column in Table 3, obtained using an independent emotion classifier model, is a key finding that highlights the significant transfer of emotional states, specifically from neutral to happy. In addition

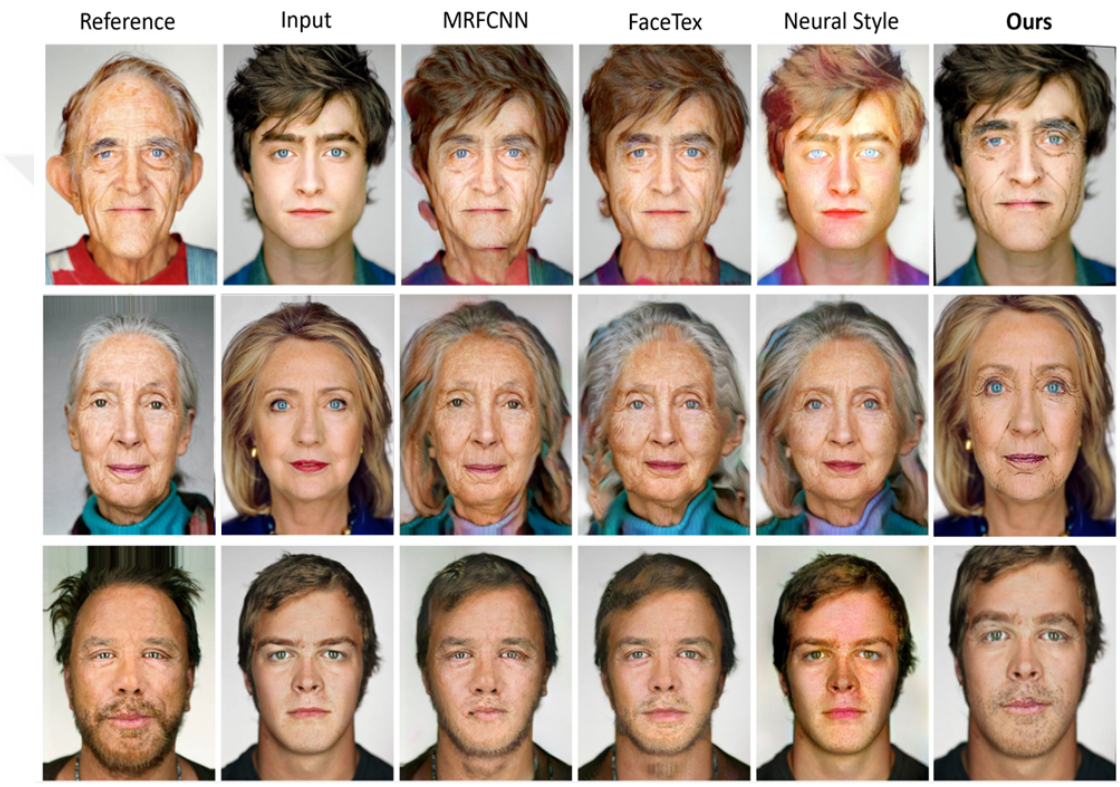

Figure 24. Visual comparison of our method with CNN-based methods (Kaur et al., 2017; Gatys et al., 2016; Li and Wand, 2016) using dataset from (Shih et al., 2014)

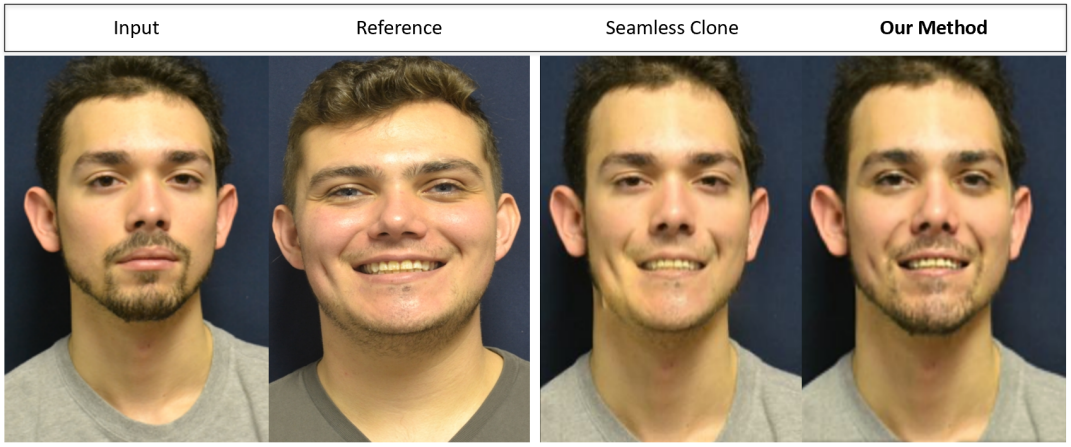

Figure 25. Comparing our method with Poisson Image Editing (Pérez et al., 2003)

to the results mentioned above, successful cases where the presented method, is used solely for texture transfer are shown in Fig. 24. This visual comparison table was directly taken from the publication titled 'Photo-realistic Facial Texture Transfer,' (Kaur et al., 2017) after which the outputs of our method were added to the far-right column. As seen in Fig. 24, our method has produced much more stable outputs with less distortion compared to CNN-based human face texture transfer approaches. Our method successfully delivers outputs by only transferring the relevant features (wrinkles and certain signs of aging) while preventing deviations in eye color, changes in hair color, and other distortions in facial features. These results visually demonstrate that our method operates effectively and robustly. In addition to the results mentioned above, in Fig. 25 an example simulation using the Poisson Image Editing method (Pérez et al., 2003) has observed that the emotional state is transferred, but the person's identity changes. Moreover, due to the nature of the cartoon-texture separation process, some uncontrollable features have been transferred within the scope of experimental studies. These results also demonstrate the method's success in texture transfer. The outputs of the presented method include uncontrollable yet potentially useful results. Figure 26 shows some texture transfer results such as beards, mustaches, glasses, ethnic traits, and aging, using the TUFTs Emotion Dataset (Panetta et al., 2020) and the Head Portrait Dataset (Shih et al., 2014).

In Fig. 27, some image pairs from the TUFTs Emotion Dataset (Panetta et al., 2020) are used as inputs to a model operating with a latent space mapping approach,

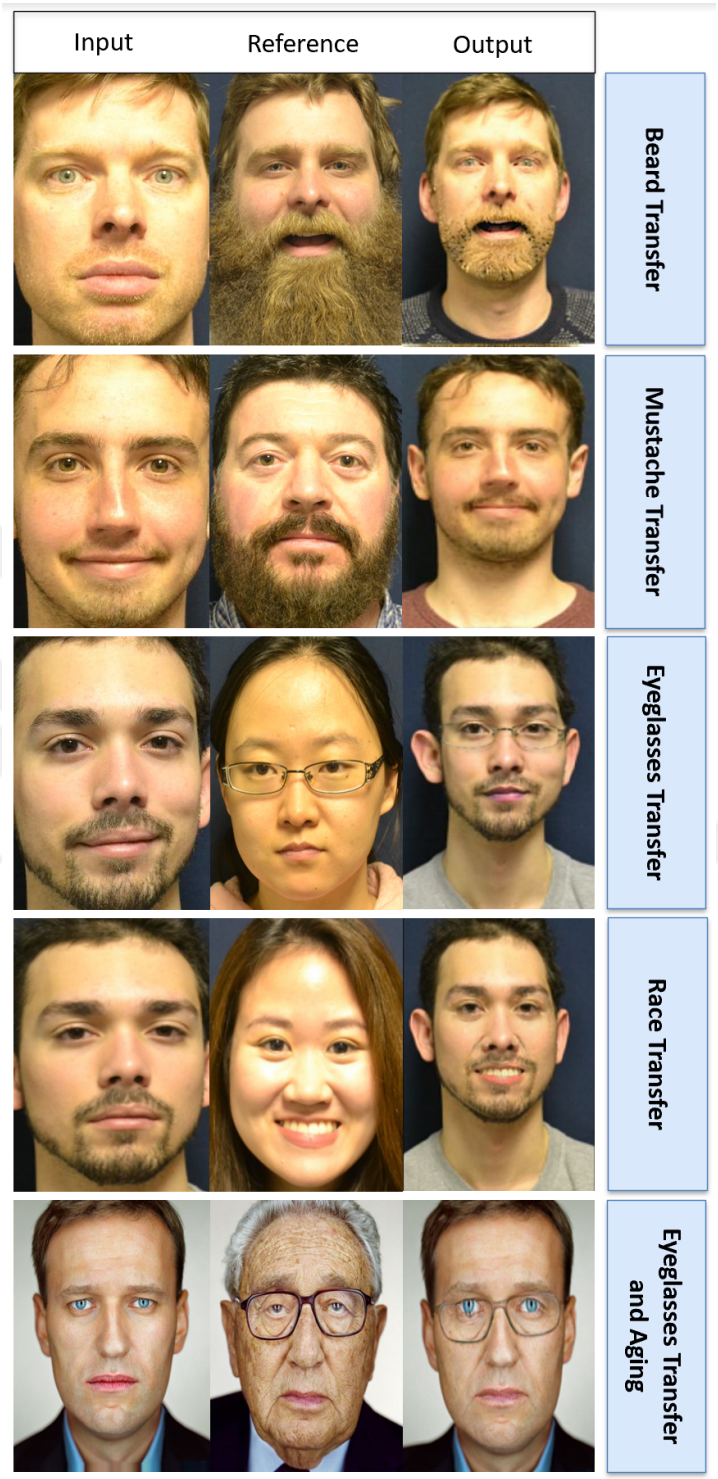

Figure 26. Results Obtained Due to the Logic of the Method, Which Are Uncontrollable

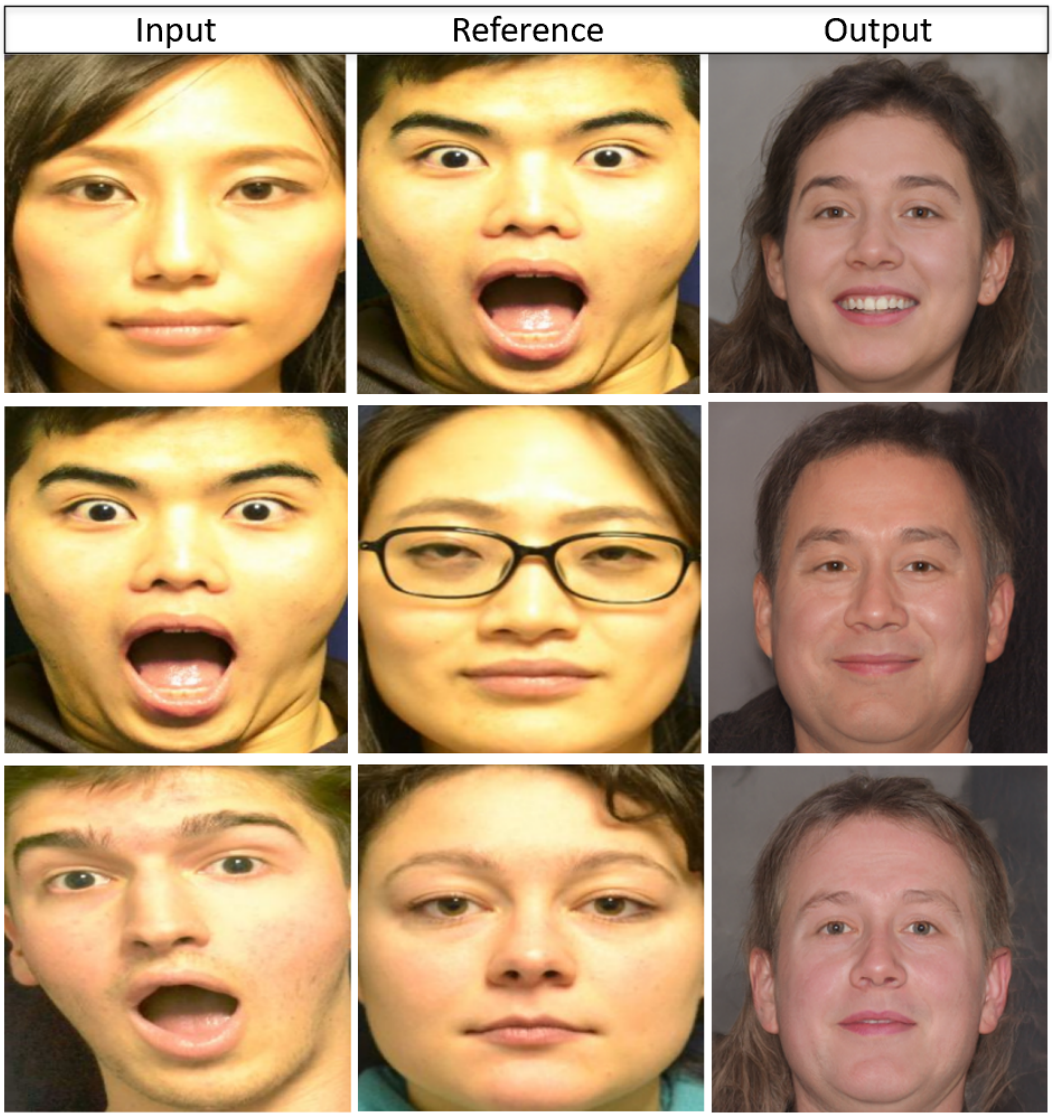

Figure 27. Output of the latent space mapping method (Nitzan et al., 2020) with inputs from TUFTs Emotion Dataset

and the resulting outputs are displayed. It is important to note that generative AI models produce outputs based on the probability distribution function of the dataset they are trained on; hence, when images in styles not present in the dataset are inputted, the outputs bear traces of the training data. In Fig. 27, when we examine the input and reference pairs, we see that the latent space mapping-based model (Nitzan et al., 2020) preserves fundamental features like gender in the input image while only transferring the emotional state. However, the identity information in the input image is completely lost. The feature of our proposed method not being dependent on a dataset demonstrates its adaptability and its stable performance in adapting to images of different sizes and styles, which is where our method excels compared to advanced GAN-based models.

4.2.1. More Results with Emotions

The following figures display various outputs from different simulations, along with the results of emotion classification. In Fig. 28, 29, 30, and 31, a green frame around the column of the dominant emotion indicates that the reference image's emotional state has been successfully transferred to the input image, while a red frame indicates that the reference emotional state was not made dominant in the input image.

Figures 28, 29, and 30 display visual outputs from various simulations conducted using the TUFTs Emotion Dataset (Panetta et al., 2020) and illustrate the dominant emotional states classified by an independent classifier (Ryumina et al., 2022). Figure 31 is prepared in the same format but includes simulation results using the RafD (Langner et al., 2010) dataset. In all four figures, the emotional states of the input and reference images, as specified in their respective datasets, are noted on the left side of the images. As shown in these graphics, transitions from emotions like "Surprise" or "Happy" to "Neutral" pose a challenge. Even though the visual manipulation is optimally executed, the emotion classification model does not solely consider the mouth and eye areas; eyebrows, forehead lines, and dimples also play a significant role in determining emotional states.

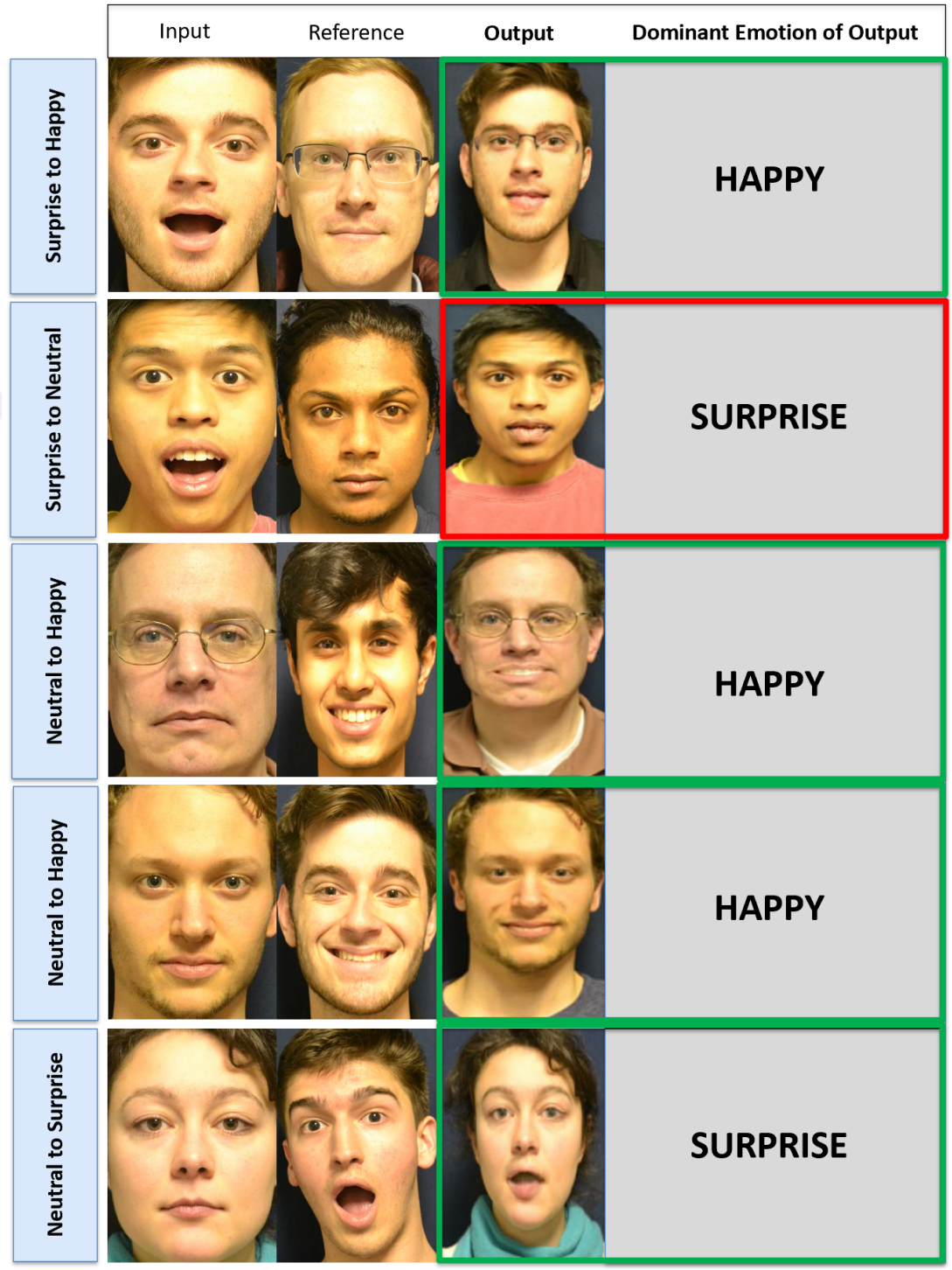

Figure 28. Visual outputs with classification result on TUFTs Emotion Dataset

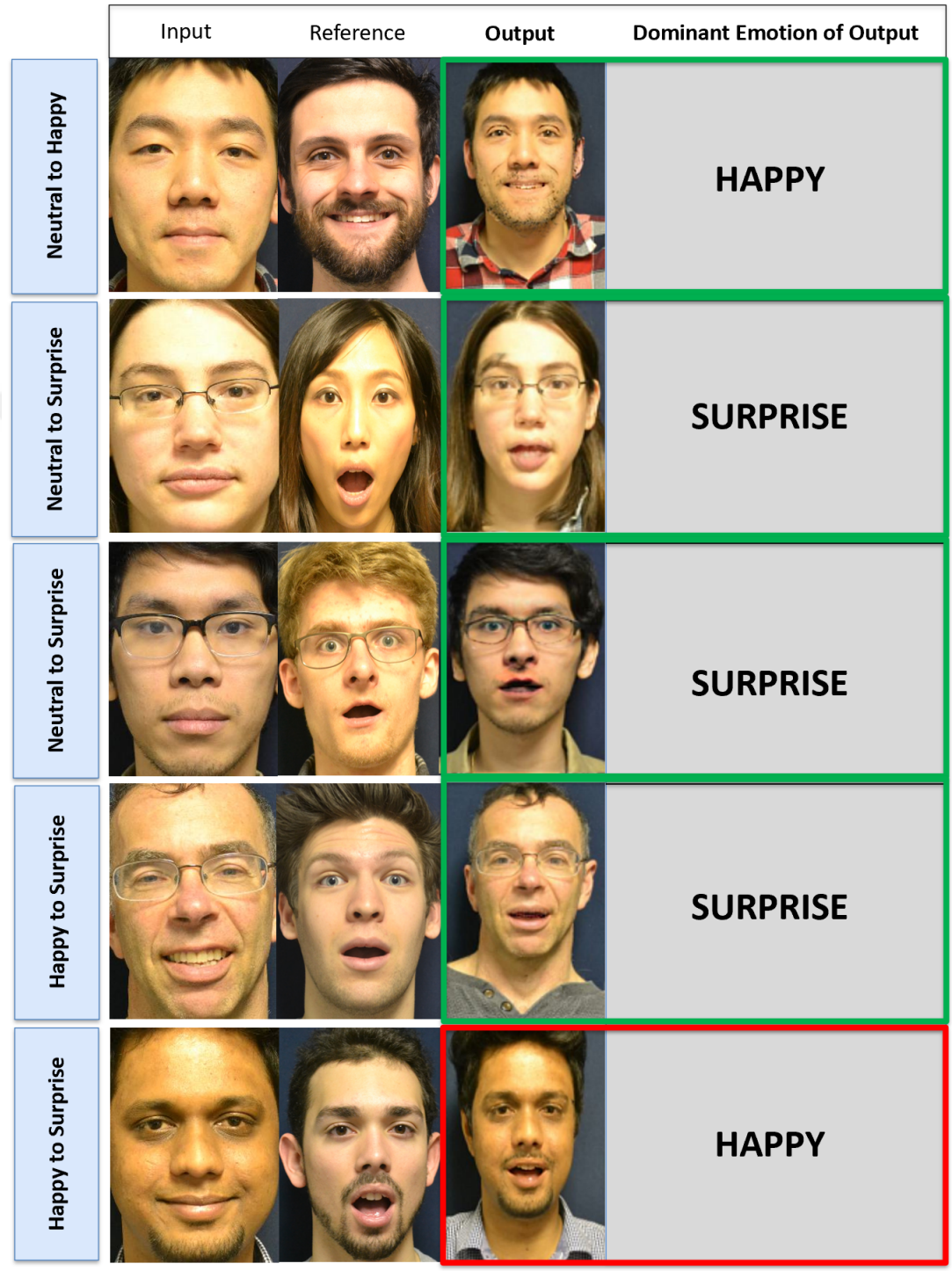

Figure 29. Visual outputs with classification result on TUFTs Emotion Dataset

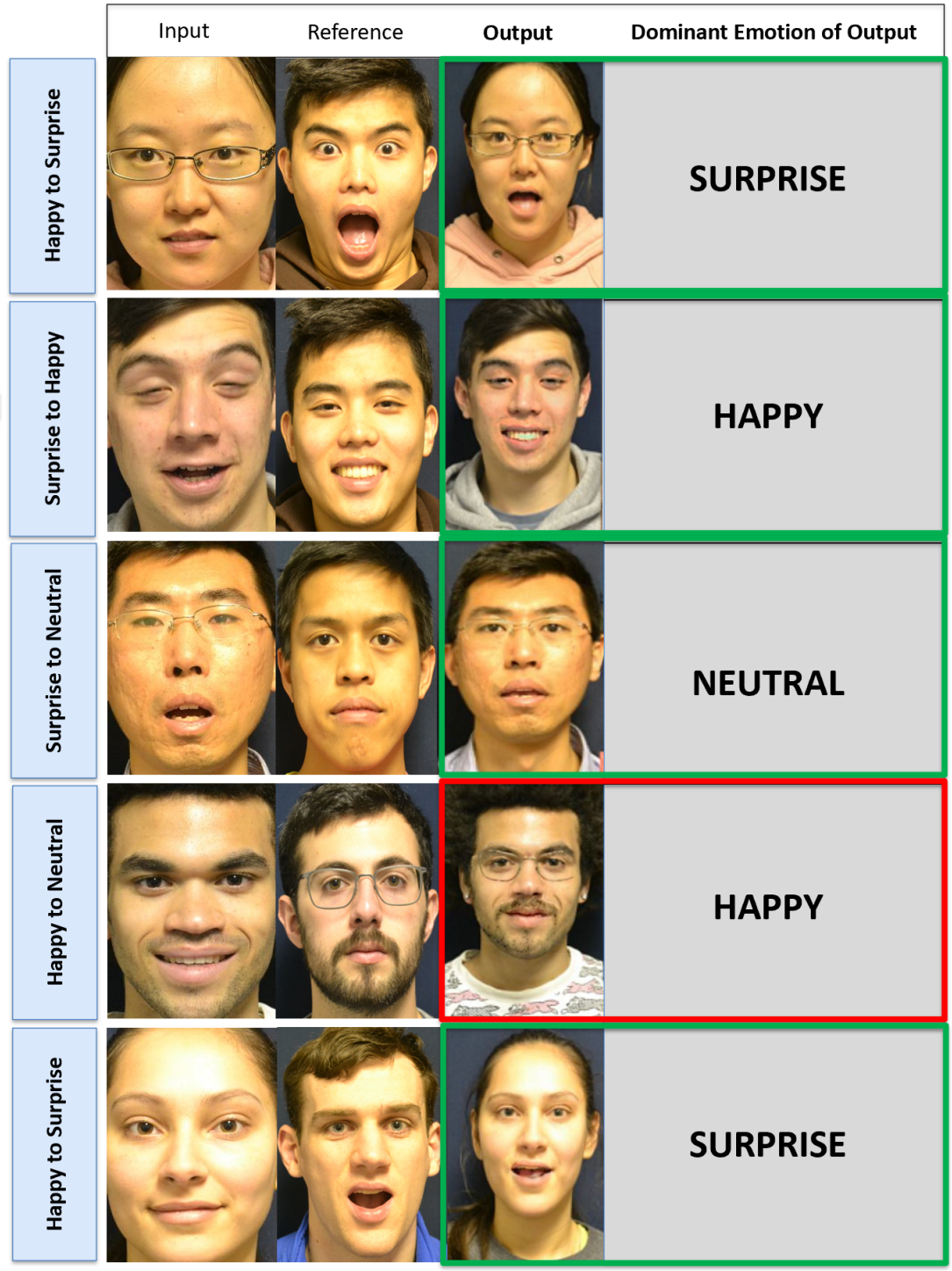

Figure 30. Visual outputs with classification result on TUFTs Emotion Dataset

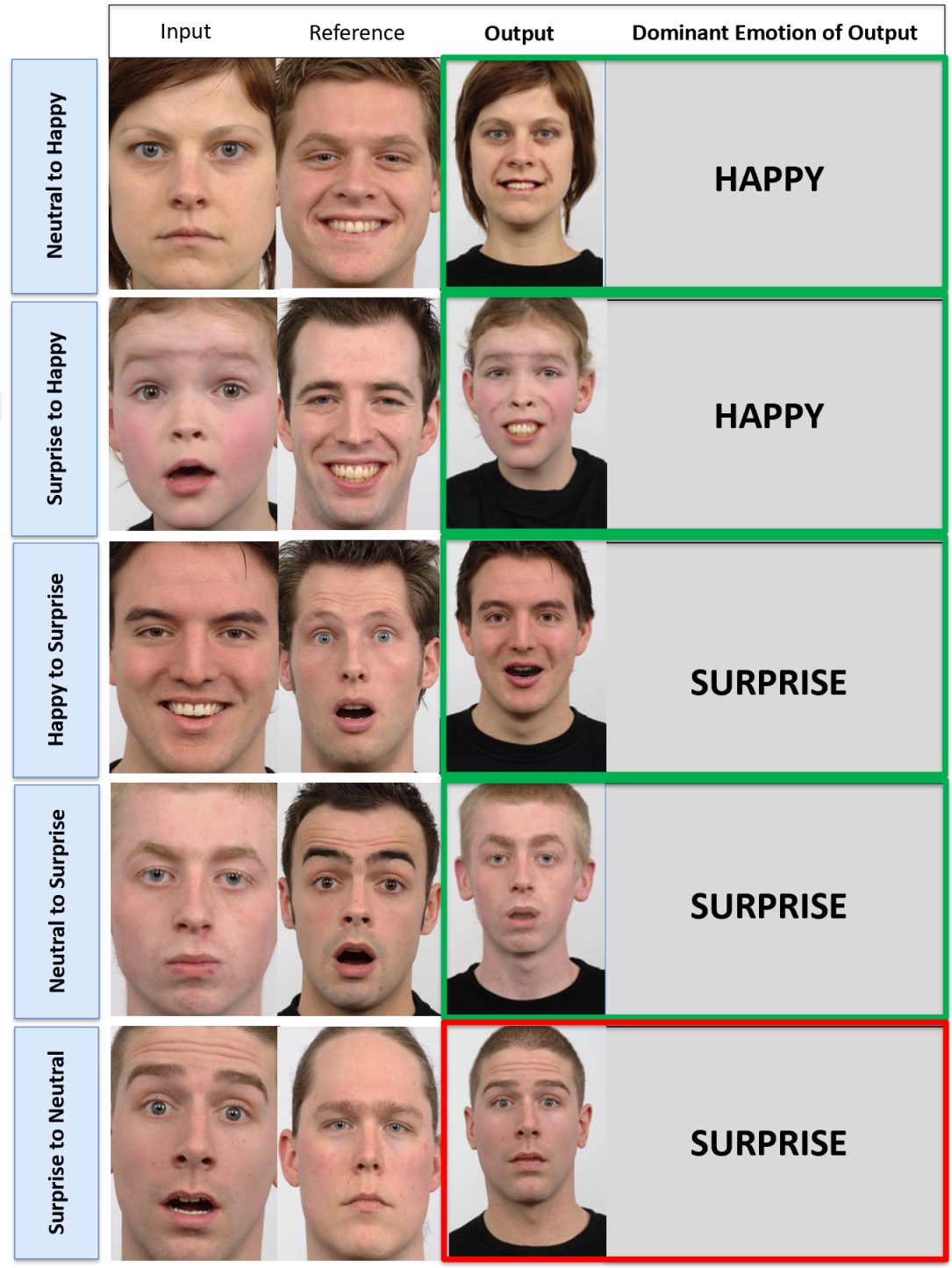

Figure 31. Visual outputs with classification result on RafD Dataset

CHAPTER 5: CONCLUSIONS

The method we have presented is designed to generalize the problem of facial texture transfer and emotion transfer. The method produces results comparable to the outputs of generative deep networks designed for similar purposes. Although the method does not achieve as successful results in emotion transfer as generative networks, it significantly increases the probability of the target emotional state in a deep learning-based emotion state classifier. Particularly, the ability of the method to be practically applied to images of any size, and its success in the FID and PSNR metrics, highlight the method presented in this article. Additionally, the method demonstrates that successful results can be achieved in the problem of texture and emotion transfer by integrating features obtained from sparse representation through an optimization process. The method allows for transferring features such as wrinkles, dimples, and the mouth and attributes like glasses and beards.

The low hardware requirements, fast performance, independence from the dataset, and the absence of a model training process, combined with the visual results being acceptable according to the presented metrics, demonstrate that the proposed method can be preferred as a robust approach. By improving the proposed method, a higher percentage of emotion state transfer can be achieved, and the desired facial features to be transferred can be classified, evolving into a more flexible transfer method.

Additionally, the method, with its included combined loss function and multi-scale optimization approach, is conducive to the emergence of new derivatives that will allow for partial cloning between images containing different objects, not limited to human faces.

REFERENCES

Abiantun, R., Juefei-Xu, F., Prabhu, U. and Savvides, M. (2019) *Ssr2: Sparse signal recovery for single-image super-resolution on faces with extreme low resolutions*, Pattern Recognition, Vol. 90, pp. 308–324.

Benning, M., Möller, M., Nossek, R. Z., Burger, M., Cremers, D., Gilboa, G. and Schönlieb, C.-B. (2017) Nonlinear spectral image fusion, *in* F. Lauze, Y. Dong and A. B. Dahl (eds), *Scale Space and Variational Methods in Computer Vision*, Cham, Springer International Publishing, pp. 41–53.

Cai, T. T. and Wang, L. (2011) *Orthogonal matching pursuit for sparse signal recovery with noise*, IEEE Transactions on Information Theory, Vol. 57 (7), pp. 4680–4688.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S. and Choo, J. (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah ,18-22 June, 2018*.

Choi, Y., Uh, Y., Yoo, J. and Ha, J. (2020) *Stargan v2: Diverse image synthesis for multiple domains*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. pp. 8185–8194. Publisher Copyright: © 2020 IEEE.; 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020 ; Conference date: 14-06-2020 Through 19-06-2020.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I. and Zafeiriou, S. (2019) *Retinaface: Single-stage dense face localisation in the wild*, CoRR, Vol. abs/1905.00641.

Elad, M., Starck, J.-L., Querre, P. and Donoho, D. (2005) *Simultaneous cartoon and texture image inpainting using morphological component analysis (mca)*, Applied and Computational Harmonic Analysis, Vol. 19 (3), pp. 340–358. Computational Harmonic Analysis - Part 1.

Frigo, O., Sabater, N., Delon, J. and Hellier, P. (2016) Split and match: Example-based adaptive patch sampling for unsupervised style transfer, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 553–561.

Garrido, P., Valgaerts, L., Rehmsen, O., Thormaehlen, T., Perez, P. and Theobalt, C.

(2014) Automatic face reenactment, *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE.

Gatys, L. A., Ecker, A. S. and Bethge, M. (2016) Image style transfer using convolutional neural networks, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C. and Bengio, Y. (2014) Generative adversarial nets, *Neural Information Processing Systems*.

Guo, X., Kan, M., He, Z., Song, X. and Shan, S. (2021) *Image style disentangling for instance-level facial attribute transfer*, Computer Vision and Image Understanding, Vol. 207, pp. 103205.

He, K., Zhang, X., Ren, S. and Sun, J. (2015) *Deep residual learning for image recognition*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. pp. 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, *Neural Information Processing Systems*.

Ho, J., Jain, A. and Abbeel, P. (2020) *Denoising diffusion probabilistic models*, ArXiv, Vol. abs/2006.11239.

Horé, A. and Ziou, D. (2010) Image quality metrics: Psnr vs. ssim, *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369.

Huang, H., Lin, Y., Song, M., Bu, J. and Chen, C. (2010) Face aging by sparse representation, *in* G. Qiu, K. M. Lam, H. Kiya, X.-Y. Xue, C.-C. J. Kuo and M. S. Lew (eds), *Advances in Multimedia Information Processing - PCM 2010*, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 571–582.

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A. (2016) *Image-to-image translation with conditional adversarial networks*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. pp. 5967–5976.

Karras, T., Laine, S. and Aila, T. (2018) *A style-based generator architecture for generative adversarial networks*, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vol. pp. 4396–4405.

Kaur, P., Zhang, H. and Dana, K. J. (2017) *Photo-realistic facial texture transfer*, 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Vol. pp. 2097–2105.

Kazemi, V. and Sullivan, J. (2014) *One millisecond face alignment with an ensemble of regression trees*, 2014 IEEE Conference on Computer Vision and Pattern Recognition, Vol. pp. 1867–1874.

Kingma, D. P. and Welling, M. (2019) *An introduction to variational autoencoders*, Foundations and Trends® in Machine Learning, Vol. 12 (4), pp. 307–392.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks, *in* F. Pereira, C. Burges, L. Bottou and K. Weinberger (eds), *Advances in Neural Information Processing Systems*, Vol. 25, Curran Associates, Inc.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T. and van Knippenberg, A. (2010) *Presentation and validation of the radboud faces database*, Cognition and Emotion, Vol. 24 (8), pp. 1377–1388.

Lawson, C. L. (1972) *Transforming triangulations*, Discrete Mathematics, Vol. 3 (4), pp. 365–372.

Li, C. and Wand, M. (2016) Precomputed real-time texture synthesis with markovian generative adversarial networks, *European Conference on Computer Vision*.

Ma, F., Xue, H., Wang, G., Zhou, Y., Rao, F., Yan, S., Zhang, Y., Wu, S., Shou, M. Z. and Sun, X. (2024) *Multi-modal generative embedding model*, ArXiv, Vol. abs/2405.19333.

Mollahosseini, A., Hasani, B. and Mahoor, M. H. (2019) *Affectnet: A database for facial expression, valence, and arousal computing in the wild*, IEEE Transactions on Affective Computing, Vol. 10 (1), pp. 18–31.

Nitzan, Y., Bermano, A. H., Li, Y. and Cohen-Or, D. (2020) *Face identity disentanglement via latent space mapping*, ACM Transactions on Graphics (TOG), Vol. 39, pp. 1 – 14.

Panetta, K., Wan, Q., Agaian, S., Rajeev, S., Kamath, S., Rajendran, R., Rao, S. P., Kaszowska, A., Taylor, H. A., Samani, A. and Yuan, X. (2020) *A comprehensive database for benchmarking imaging systems*, IEEE Transactions on Pattern Analysis

and Machine Intelligence, Vol. 42 (3), pp. 509–520.

Pérez, P., Gangnet, M. and Blake, A. (2003) *Poisson image editing*, ACM Transactions on Graphics (TOG), Vol. 22 (3), pp. 313–318.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A. and Moreno-Noguer, F. (2020) *GANimation: One-shot anatomically consistent facial animation*, International Journal of Computer Vision, Vol. 128 (3), pp. 698–713.

Radford, A., Metz, L. and Chintala, S. (2015) *Unsupervised representation learning with deep convolutional generative adversarial networks*, CoRR, Vol. abs/1511.06434.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I. (2021) *Zero-shot text-to-image generation*, ArXiv, Vol. abs/2102.12092.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2021) *High-resolution image synthesis with latent diffusion models*, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vol. pp. 10674–10685.

Rudin, L. I., Osher, S. and Fatemi, E. (1992) *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena, Vol. 60 (1), pp. 259–268.

Ryumina, E., Dresvyanskiy, D. and Karpov, A. (2022) *In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study*, Neurocomputing, Vol. .

Shao, Z., Zhu, H., Tang, J., Lu, X. and Ma, L. (2021) *Explicit facial expression transfer via fine-grained representations*, IEEE Transactions on Image Processing, Vol. 30, pp. 4610–4621.

Shih, Y., Paris, S., Barnes, C., Freeman, W. T. and Durand, F. (2014) *Style transfer for headshot portraits*, ACM Trans. Graph., Vol. 33 (4).

Sikka, H. D. (2020) *A deeper look at the unsupervised learning of disentangled representations in β-vae from the perspective of core object recognition*, ArXiv, Vol. abs/2005.07114.

Sohl-Dickstein, J. N., Weiss, E. A., Maheswaranathan, N. and Ganguli, S. (2015) *Deep unsupervised learning using nonequilibrium thermodynamics*, ArXiv, Vol. abs/1503.03585.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2015) *Rethinking the inception architecture for computer vision*, 2016 IEEE Conference on Computer

Vision and Pattern Recognition (CVPR), Vol. pp. 2818–2826.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C. and Nießner, M. (2016) *Face2face: Real-time face capture and reenactment of rgb videos*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. pp. 2387–2395.

Tibshirani, R. (1996) *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological), Vol. 58 (1), pp. 267–288.

Wu, R., Zhang, G., Lu, S. and Chen, T. (2020) *Cascade ef-gan: Progressive facial expression editing with local focuses*, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vol. pp. 5020–5029.

Yang, B. and Li, S. (2012) *Pixel-level image fusion with simultaneous orthogonal matching pursuit*, Information Fusion, Vol. 13 (1), pp. 10–19.

Yang, H., Huang, D., Wang, Y., Wang, H. and Tang, Y. (2016) *Face aging effect simulation using hidden factor analysis joint sparse representation*, IEEE Transactions on Image Processing, Vol. 25 (6), pp. 2493–2507.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E. and Wang, O. (2018) *The unreasonable effectiveness of deep features as a perceptual metric*, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vol. pp. 586–595.

Zhou, S., Xiao, T., Yang, Y., Feng, D., He, Q. and He, W. (2017) *Genegan: Learning object transfiguration and attribute subspace from unpaired data*, ArXiv, Vol. abs/1705.04932.

Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A. (2017) *Unpaired image-to-image translation using cycle-consistent adversarial networks*, 2017 IEEE International Conference on Computer Vision (ICCV), Vol. pp. 2242–2251.