ISTANBUL TECHNICAL UNIVERSITY
FACULTY OF COMPUTER AND INFORMATICS

SYNTHETIC DATA GENERATION FOR FACIAL EXPRESSION RECOGNITION

Graduation Project Final Report

Ömer Yıldırım
150190115

Department: Computer Engineering
Division: Computer Engineering

Advisor: Prof. Dr. Hazım Kemal Ekenel

June 2024

Statement of Authenticity

I hereby declare that in this study

1. all the content influenced by external references are cited clearly and in detail,
2. and all the remaining sections, especially the theoretical studies and implemented software that constitute the fundamental essence of this study is originated by my individual authenticity.

İstanbul, June 2024

Ömer Yıldırım

Acknowledgments

I extend my sincere gratitude to Professor Dr. Hazım Kemal Ekenel, my advisor, for affording me the invaluable opportunity to collaborate with him and his reputable research group, Smart Interaction and Machine Intelligence Technologies (SIMIT), on the subject of "Synthetic Data Generation for Facial Expression Recognition". His expert guidance and unwavering support were instrumental in the successful execution of this project.

Special appreciation is also extended to the esteemed researchers Ali Azmoudeh and Erdi Sarıtaş for their noteworthy support and understanding during the implementation of this precious project, which significantly contributed to the overall progress and achievement of the undertaking.

SYNTHETIC DATA GENERATION FOR FACIAL EXPRESSION RECOGNITION

(SUMMARY)

The project delves into the innovative application of Synthetic Data Generation to enhance Facial Expression Recognition (FER). The primary objectives are multi-faceted and ambitious, aimed at revolutionizing the current approaches to FER by leveraging synthetic data. This summary provides an overview of the project's goals, methodologies, datasets, and the significant advancements achieved throughout the research.

The project aims to develop a robust system capable of producing high-quality synthetic facial expression data. This system is intended to complement existing datasets, thereby mitigating the limitations and biases associated with traditional training data. By integrating hybrid datasets that combine real and synthetic data, the project seeks to enhance the accuracy and generalizability of FER models. This hybrid approach addresses the challenge of insufficient training data and aims to elevate the overall performance of these models.

The project extensively explores Generative Adversarial Networks (GANs), focusing on their ability to generate realistic synthetic facial images. The efficacy of GANs is evaluated based on their impact on FER model performance, with particular attention to generating data based on action units frequently used in FER problems. A significant objective is to contribute to the broader field of FER research. By demonstrating the advantages of synthetic data, the project addresses existing challenges and aims to propel advancements in FER methodologies.

The project utilizes GANmut, a GAN-based framework that learns an expressive and interpretable conditional space of emotions, chosen after an extensive literature review of similar GAN models such as ExprGAN and StarGAN. Through this framework, three datasets were created: SynFace15F, a synthetic dataset of facial expressions; SynFace15V, an improved version of SynFace15F, addressing domain-specific issues; and RafSynFace30V, a hybrid dataset combining the RAF-DB dataset with SynFace15V. The primary purpose of these datasets is to compare the performance of FER models trained on synthetic data with those trained on real data. The desired outcome is to replace the time-consuming and costly process of collecting and annotating natural data with the generation of synthetic data for training models.

The project trained the POSTER model, a pyramid cross-fusion transformer network for FER, using the four datasets (RAF-DB, SynFace15F, SynFace15V, and RafSynFace30V). The POSTER model, known for its state-of-the-art classification performance, was evaluated to determine the feasibility of using synthetic data in place of natural data. The results demonstrated that models trained on synthetic data exhibited comparable or superior accuracy to those trained on conventional datasets, with a target accuracy of better than the classification accuracy of models trained with natural datasets.

Significant improvements were made to the GANmut framework, including color transfer in LAB color space to ensure the skin color of the synthetic image matches the original image and the alpha blending technique to achieve a smooth transition in the paste operation after emotion editing. These enhancements resulted in better test results by addressing domain-specific issues, underscoring the importance of the generation process in the overall performance of FER models.

The project has made several key contributions to FER research. The successful training and testing of FER models with synthetic data highlight its potential as a substitute for natural data. The combination of synthetic and real data in hybrid datasets enhances model classification performance by increasing data diversity and quantity. The analysis of models trained on different versions of synthetic datasets underscores the critical role of the generation process in model performance. The generation of a dataset containing 15,000 images, SynFace15V, provides a valuable resource for various FER applications.

This project successfully demonstrates the potential of synthetic data to replace natural datasets in FER model training. The comprehensive evaluation of FER models trained on synthetic and hybrid datasets reveals significant improvements in performance, driven by increased data diversity and optimized generation processes. By advancing synthetic data generation techniques and applying them to FER, this project contributes valuable insights and resources to the field, paving the way for future research and applications in facial expression recognition.

YÜZ İFADESİ TANIMA İÇİN SENTETİK VERİ ÜRETİMİ

(ÖZET)

Bu proje, Sentetik Veri Üretiminin Yüz İfadesi Tanıma (FER) süreçlerini geliştirmek amacıyla uygulamasını ele almaktadır. Projenin ana hedefleri çok yönlü ve iddialıdır; mevcut FER yaklaşımlarını sentetik veri kullanarak devrim niteliğinde iyileştirmeyi amaçlamaktadır. Bu özet, projenin hedefleri, metodolojileri, veri kümeleri ve araştırma boyunca elde edilen önemli ilerlemeler hakkında genel bir bakış sunar.

Proje, yüksek kaliteli sentetik yüz ifadesi verileri üretebilen yüksek performans bir sistem geliştirmeyi amaçlamaktadır. Bu sistem, mevcut veri kümelerini tamamlamak amacıyla tasarlanmıştır ve geleneksel eğitim verileriyle ilgili sınırlamaları ve bu eğitim verilerinde yer alan doğal yanlılılıkları azaltmayı hedefler. Gerçek ve sentetik verileri birleştiren hibrit veri kümelerini entegre ederek, proje FER modellerinin sınıflandırma doğruluğunu ve genellenebilirliğini artırmayı hedefler. Bu hibrit yaklaşım, yetersiz eğitim verileri sorununu ele alır ve bu modellerin genel performansını yükseltmeyi amaçlar.

Proje, Generative Adversarial Networks (GAN'lar) kullanılarak gerçekçi sentetik yüz görüntüleri üretme yeteneğini kapsamlı bir şekilde incelemektedir. GAN'ların etkinliği, FER modeli performansı üzerindeki etkileri temel alınarak değerlendirilmektedir ve FER sorunlarında yaygın olarak kullanılan eylem birimlerine dayalı veri üretimine özellikle dikkat edilmektedir. Sentetik verilerin avantajlarını göstererek, proje mevcut zorlukları ele alır ve FER metodolojilerindeki ilerlemeleri teşvik etmeyi amaçlar.

Proje, geniş bir literatür incelemesinin ardından ExprGAN ve StarGAN gibi benzer GAN modellerinin araştırılması sonucunda seçilen, duyguların ifade edilebilir ve yorumlanabilir bir koşullu alanını öğrenen GANmut adlı GAN tabanlı bir çerçeve kullanmaktadır. Bu çerçeve aracılığıyla üç veri kümesi oluşturulmuştur: Sentetik yüz ifadeleri veri kümesi olan SynFace15F; SynFace15F'in iyileştirilmiş ve domain (alan) spesifik sorunlarını ele alan versiyonu olan SynFace15V; ve RAF-DB veri kümesi ile SynFace15V'yi birleştiren hibrit bir veri kümesi olan RafSynFace30V. Bu veri kümelerinin ana amacı, sentetik verilerle eğitilen FER modellerinin performansını, RAF-DB datasetinde yer alan gerçek verilerle eğitilenlerle karşılaştırmaktır. Arzu edilen sonuç, modelleri eğitmek için doğal verilerin toplanması ve anotasyon yapılması sürecinin zaman alıcı ve maliyetli işlemlerinin yerine, sentetik verilerin üretimiyle bu süreçleri ikame etmektir.

Proje, POSTER modelini (FER için bir piramit çapraz füzyon transformatör ağı) RAF-DB, SynFace15F, SynFace15V ve RafSynFace30V olmak üzere dört veri kümesi kullanarak eğitmiştir. Sınıflandırma performansı açısından FER araştırmalarında en son teknoloji ürünü olan POSTER modeli, sentetik verilerin doğal verilerin yerine kullanılabilirliğini belirlemek amacıyla değerlendirilmiştir. Sonuçlar, sentetik verilerle eğitilen modellerin, geleneksel veri kümeleriyle eğitilen modellere kıyasla benzer veya üstün doğruluk sergilediğini göstermiştir.

GANmut çerçevesine, sentetik görüntünün cilt renginin orijinal görüntüyle eşleşmesini sağlamak için LAB renk uzayında renk aktarımı ve duygu düzenlemesi sonrası yapıştırma işleminin düzgün geçişini sağlamak için alfa harmanlama tekniği gibi önemli iyileştirmeler yapılmıştır. Bu iyileştirmeler, alan özel sorunları ele alarak daha iyi test sonuçları elde edilmesini sağlamış ve sentetik veri üretim sürecinin FER modellerinin genel performansındaki önemini vurgulamıştır.

Proje, FER araştırmalarına birkaç önemli katkı sağlamıştır. Sentetik verilerle eğitilen FER modellerinin başarılı bir şekilde eğitilmesi ve test edilmesi, bu verilerin doğal verilerin yerine geçme potansiyelini vurgulamaktadır. Hibrit veri kümelerinde sentetik ve gerçek verilerin birleştirilmesi, ver, çeşitliliği ve miktarını artırarak model sınıflandırma performansını arttırmaktadır. Farklı sentetik veri kümeleriyle eğitilen modellerin analizi, sentetik veri üretim sürecinin model performansı üzerindeki kritik rolünü ortaya koymaktadır. Sentetik veri kümesi olan SynFace15V'nin 15.000 görüntü içeren veri kümesinin üretilmesi, çeşitli FER uygulamaları için değerli bir veri seti olarak kaynak sağlamaktadır.

Bu proje, sentetik verilerin FER model eğitiminde doğal veri kümelerinin yerine kullanılma potansiyelini başarıyla göstermektedir. Sentetik ve hibrit veri kümeleriyle eğitilen FER modellerinin kapsamlı değerlendirmesi, veri çeşitliliği artırılarak ve üretim süreçleri optimize edilerek performansta önemli iyileştirmeler ortaya koymaktadır. Sentetik veri üretim tekniklerini geliştirerek ve bunları FER'e uygulayarak, proje alanına değerli bilgiler ve kaynaklar sağlayarak yüz ifadesi tanıma araştırmalarında gelecekteki araştırmalara ve uygulamalara zemin hazırlamaktadır.

Contents

1 Introduction and Project Summary

In recent years, deep learning has demonstrated remarkable efficacy in various computer vision tasks, such as image classification and face recognition. The notable success of deep learning can be attributed to the robust representation capabilities of deep neural networks and the availability of extensive labeled training data. While deep learning methods for Facial Expression Recognition (FER) have been developed, their performance in unconstrained environments remains unsatisfactory. This inadequacy is partly attributed to facial expression databases, which typically have limited training data. Despite the abundance of diverse facial images on the Internet, the manual annotation of these images is both time-consuming and costly. Consequently, training deep neural networks with a restricted number of labeled samples presents a non-trivial challenge.

To address the issue of insufficient training data, some studies have sought to leverage auxiliary data to effectively train neural networks for FER. For instance, a deep Convolutional Neural Network (CNN) architecture has been proposed and trained on hybrid databases comprising seven distinct facial expression databases. However, the bias among these databases introduces challenges such as overfitting and potentially diminishing performance on the target database. Developing an appropriate strategy for fine-tuning is demanding, given that deep networks are pre-trained with high capacity on large-scale data. Despite the application of techniques like data augmentation and dropout to mitigate overfitting, their impact on performance improvement is limited. Consequently, enhancing FER performance under conditions of insufficient training data remains an ongoing challenge.

The overarching objective of this project is to harness synthetic data generation to enrich the developmental landscape of facial expression recognition systems. The methodology involves training models on facial images that transcend the boundaries of reality, being algorithmically created by Generative Adversarial Networks (GAN) models. This innovative approach seeks to augment the volume and diversity of training data, promising enhancements in the refinement and performance of facial expression recognition models.

Recently, there has been growing interest in a novel generative model known as the Generative Adversarial Network (GAN) for its promising ability to generate high-quality data. The ascendancy of deep learning techniques in unraveling intricate probability distributions across diverse data types is a cornerstone of this endeavor. Notably, Generative Adversarial Networks (GANs), introduced by Goodfellow et al. in 2014[1], have emerged as a formidable tool in this context. Operating within a two-player min-max framework, GANs comprise a generative model and an adversary model. The generative model strives to encapsulate the underlying data distribution, while the adversary, or discriminator, discerns between authentic data and synthetic samples. The iterative interplay between these models persists until the generated samples attain indistinguishability from real data. Typically composed of a generator network and a discriminator network engaged in adversarial optimization, GAN can produce synthetic samples emulating the distribution of real samples from training data.

Notably, GAN has been successfully employed in face-related tasks such as posed face synthesis [2] and facial attributes transfer [3]. These methods generate photorealistic facial images sharing the same identity as input facial images. Leveraging the similarities between synthetic and real images, these generated images, featuring variations in conditions like expressions and poses, can serve as valuable data augmentation to address the challenge of limited training data in deep learning.

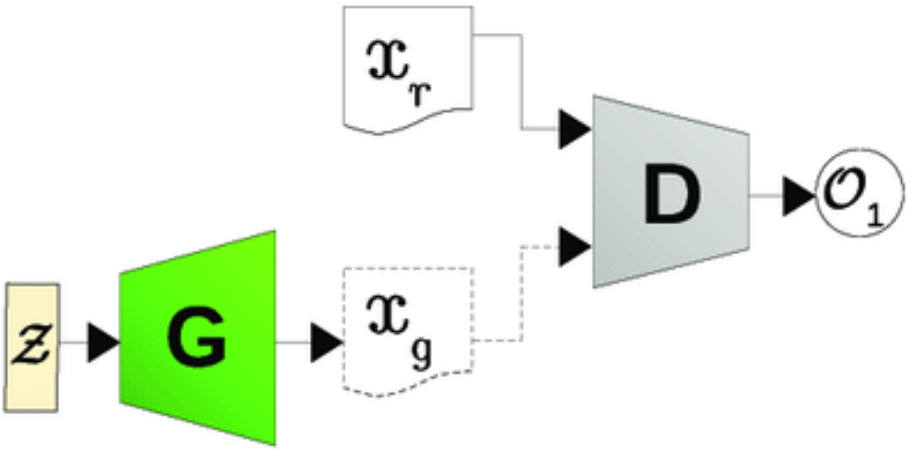

Figure 1.1: General Structure of vanilla GAN [REFERENCE] [1]; Z: input noise, G: Generator, D: Discriminator, xr: real sample, xg: generated sample, O1: Output of binary classification to real/fake.

A primary focus lies in the creation of synthetic facial images tailored for the training of FER models. This approach, driven by the rationale to supplement existing datasets, often originating from professional environments or gleaned from publicly available images, introduces a novel dimension. By integrating non-existing expressive faces generated through advanced algorithms, the project aims to diversify the training data, thereby enhancing the robustness of the models.

A pivotal purpose of the project involves a meticulous evaluation and comparison of the performance of FER models trained on conventional datasets against those trained on synthetic data. The emphasis is squarely on model classification accuracy, providing nuanced insights into the efficacy of synthetic data in bolstering overall model performance.

The comparative analysis of model performances represents a significant scientific endeavor, showcasing the potency of AI-generated data in training facial expression recognition models. Moreover, this project envisions contributing substantively to future FER research and applications by not only demonstrating the viability of synthetic data but also by providing an increased volume of diverse and meticulously crafted data.

The system to be implemented follows a multi-step approach. First, the GANmut model, a GAN-based framework [4] for emotion editing on natural images, is used to generate synthetic datasets. GANmut learns an expressive and interpretable conditional space of emotions.

For this project, after an extensive literature review of similar GAN models like ExprGAN [5] and StarGAN [6], GANmut was selected. It is particularly effective in generating realistic facial expressions that do not exist in reality, thus expanding the diversity and volume of training data.

Two versions of the synthetic dataset are generated: SynFace15F and SynFace15V. SynFace15F does not address the domain problem of mismatched color tones between synthetic and real images, nor does it employ variable emotion intensity. SynFace15V, on the other hand, includes a two-step domain adaptation process: color transfer in the LAB color space to match skin tones and an alpha blending technique to ensure smooth integration of the synthetic face into the original image. This process results in more realistic and high-quality synthetic data. Additionally, SynFace15V employs a variable emotion intensity, enhancing the data's representativeness and variability.

Next, the synthetic datasets are resized and prepared to match the resolution and features of the RAF-DB dataset, which is 100x100 pixels. This involves using MTCNN models for face detection in synthetic images to ensure that the synthetic datasets align perfectly with the natural ones. A hybrid dataset, RafSynFace30V, is also created by combining the RAF-DB dataset with SynFace15V to measure potential performance improvements when synthetic and real data are combined.

The third step involves training the POSTER model, a pyramid cross-fusion transformer network, with the four datasets: RAF-DB [7], SynFace15F, SynFace15V, and RafSynFace30V. POSTER is known for its state-of-the-art performance in FER, leveraging advanced network architectures to achieve high classification accuracy.

Finally, a comparative analysis of these models is conducted to evaluate their classification accuracy. This analysis includes testing on an independent dataset from the literature, specifically the validation dataset of AffectNet [8], which contains 500 images per emotion class (neutral, happy, angry, sad, fear, surprise, disgust). The goal is to determine whether synthetic data can effectively replace natural data in FER model training and how combining synthetic and real data in hybrid datasets affects model performance.

This project not only aims to demonstrate the potential of synthetic datasets like SynFace15V in replacing natural datasets for FER model training but also to show that hybrid datasets can significantly enhance model performance by increasing data diversity and quantity. Furthermore, the project highlights the importance of the generation process, particularly the domain adaptation steps, and variable emotion intensity, in improving the quality and effectiveness of synthetic data.

In conclusion, by addressing the challenges of limited and biased training data through synthetic data generation, this project contributes valuable insights and advancements to the FER field. The comprehensive evaluation of FER models trained with synthetic and hybrid datasets underscores the practical applications of this research, paving the way for future developments in facial expression recognition using AI-generated data.

2 Comparative Literatre Survey

Facial expression synthesis has been a subject of significant research, with several pioneering works contributing to the understanding and advancement of this field. Among these, the study by Susskind et al. [9] stands out for its incorporation of constraints like 'raised eyebrows' on generated facial samples. The framework utilizes a Deep Belief Network (DBN) with two hidden layers of 500 units, incorporating the Facial Action Coding System (FACS) vector and identity to generate faces with diverse expressions. Subsequently, the emergence of Generative Adversarial Network (GAN) models introduced innovative approaches to facial expression synthesis.

DyadGAN [10] and ExprGAN [5] are two such models designed explicitly for face generation. DyadGAN focuses on generating facial images conditioned on the expressions of a dyadic conversation partner, while ExprGAN overcomes challenges related to controlling expression intensity without relying on intensity-labeled training data. This is achieved through an expression controller module and an identity-preserving loss function.

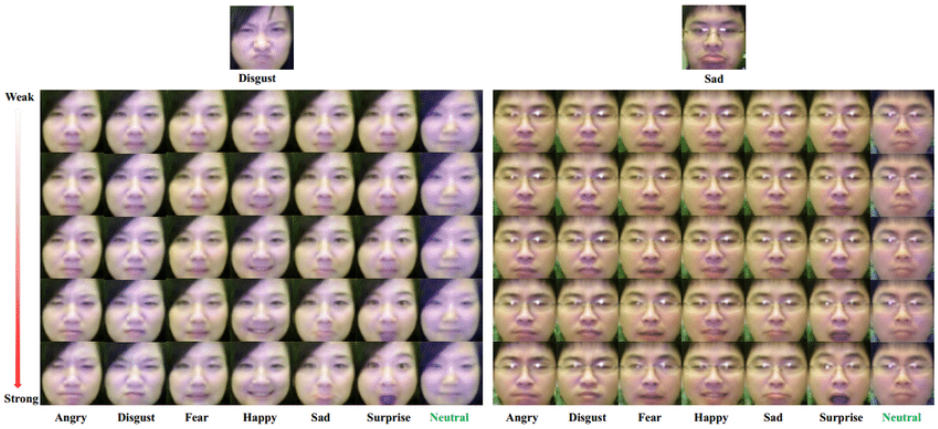

Figure 2.1: ExprGAN [5] for image generation of facial expression.

One noteworthy GAN-based model, G2GAN by Song et al. [11], offers fine-grained control over target expressions and facial attributes, providing realistic and identity-preserving images. The model leverages unpaired training with a pair of GANs—one removing the expression while the other synthesizes it. Additionally, StarGAN [6] presents a scalable solution for multi-domain image-to-image translation using a unified GAN model. Attributes like hair color, gender, and age can be modified based on the desired values.

Attribute editing GAN (AttGAN) [12] offers a framework for editing attributes among a set of face images, employing adversarial loss, reconstruction loss, and attribute classification constraints. DIAT [13], CycleGAN [14], and IcGAN [15] serve as baseline models for comparison.

In 2018, Qiao et al. [16] extended G2GAN, introducing a model based on Variational Autoencoder GANs (VAEGANs) for synthesizing facial expressions from a single image and landmarks. Unlike ExprGAN, their model does not require the target class label and dispenses with the need for a neutral expression as an intermediate level in the transfer procedure. Pumarola et al. [17] utilize facial Action Units as a one-hot vector for unsupervised expression synthesis.

The advent of Generative Adversarial Networks (GANs) revolutionized facial expression synthesis. StarGAN [3] addressed multi-domain image-to-image translation across databases by introducing a mask vector to disregard unlabelled categories during training. This innovation allows unlabelled facial images to be translated into expressions while preserving information from other domains.

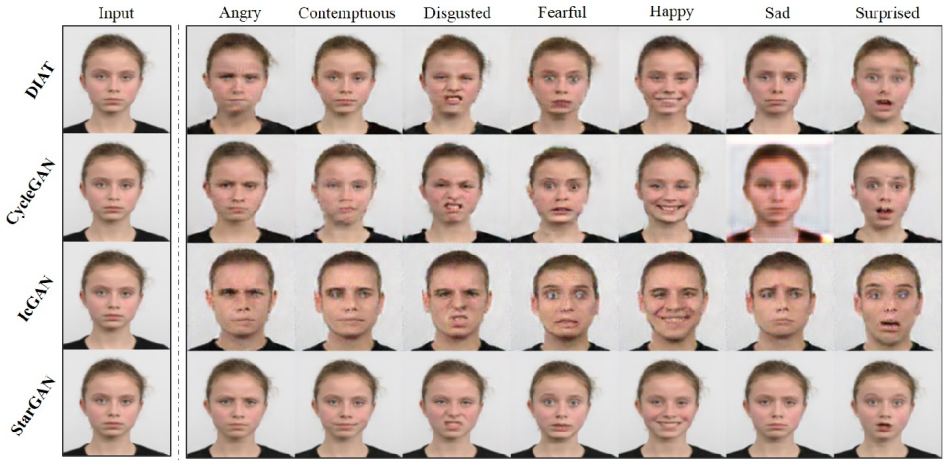

Figure 2.2: Visual comparison of the GAN models [13], [14], [15], [6] in order (top to bottom) for facial expression synthesis on RaFD dataset [18], images are in courtesy of the reviewed papers.

Additionally, StarGAN [6] presents a scalable solution for multi-domain image-to-image translation using a unified GAN model. Attributes like hair color, gender, and age can be modified based on the desired values.

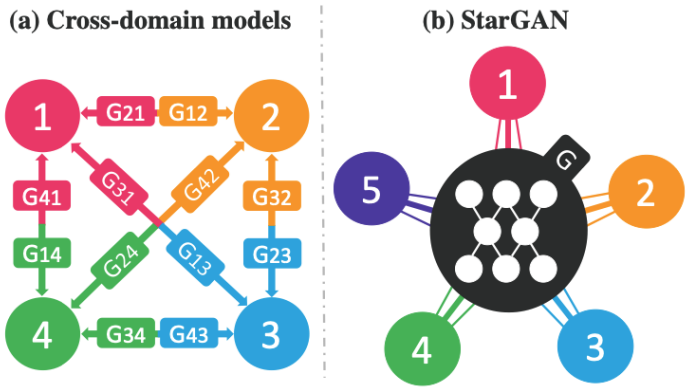

Figure 2.3: Comparison between cross-domain models and the unified model, StarGAN [6].

Furthermore, a Facial Expression Synthesis Generative Adversarial Network (FESGAN) [19] undergoes pre-training to generate facial images exhibiting diverse facial expressions. Notably, FESGAN is meticulously designed to enhance training image diversity by generating images featuring new identities derived from a predefined distribution. Subsequently, an expression recognition network undergoes joint learning with the pre-trained FESGAN within an integrated framework.

Specifically, the classification loss derived from the recognition network is employed to concurrently optimize the performance of both the recognition network and the FESGAN generator. Furthermore, to mitigate the challenge posed by data bias between real and synthetic images, this study proposes an intra-class loss incorporating a novel Real Data-Guided Back-Propagation (RDBP) algorithm. The RDBP algorithm is employed to mitigate intra-class variations among images belonging to the same class, thereby significantly enhancing the ultimate performance of the proposed approach.

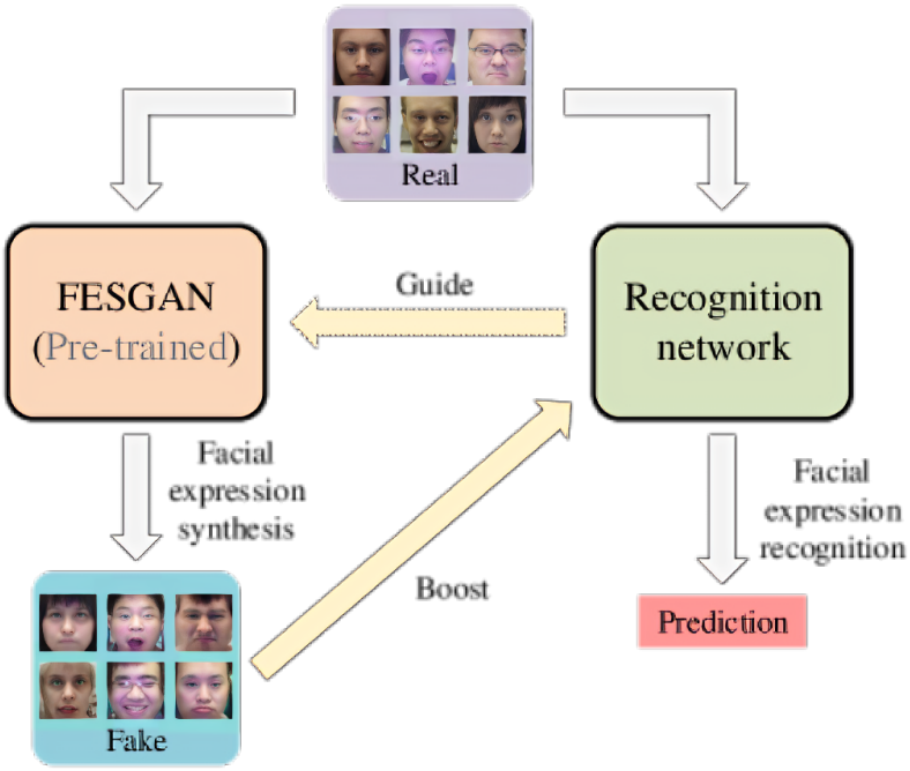

Figure 2.4: Visual comparison of the GAN models [13], [14], [15], [19] in order (top to bottom) for facial expression synthesis on RaFD [18] dataset, images are in courtesy of the reviewed papers.

The GANmut model represents a significant advancement in the realm of facial expression synthesis, aligning closely with the innovative strides made by other GAN-based models like ExprGAN, StarGAN, and G2GAN. GANmut is designed to control emotion intensity, much like ExprGAN, which incorporates an expression controller module and an identity-preserving loss function to manage expression intensity without relying on intensity-labeled training data. However, GANmut extends this capability by learning an expressive and interpretable conditional space of emotions, allowing for a more nuanced control of expression intensity.

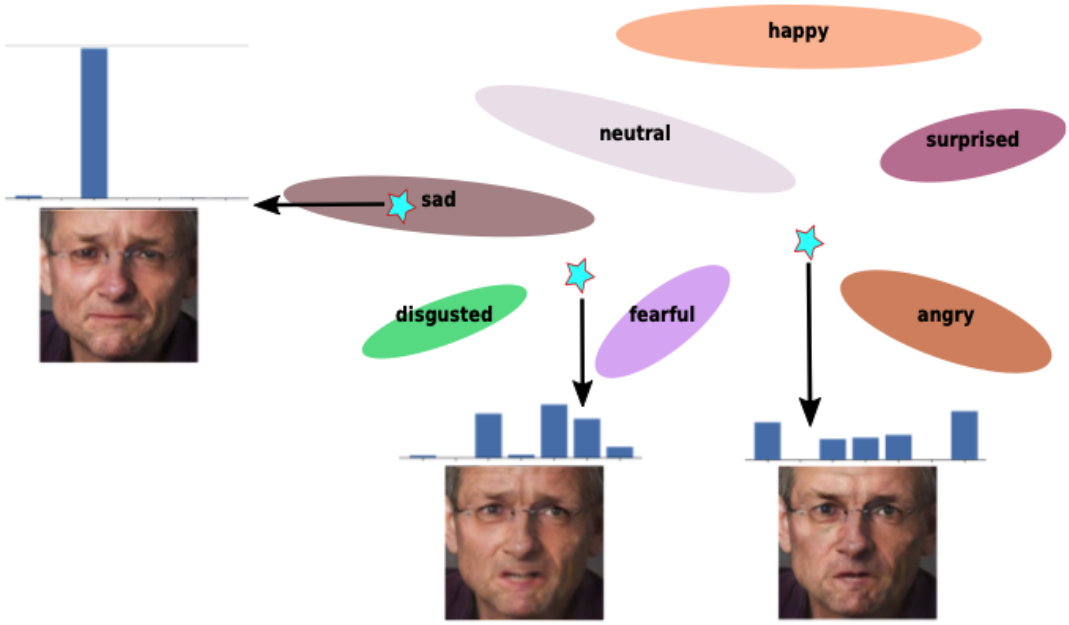

Figure 2.5: Learned conditional space for the Gaussian model of GANmut (GGANmut). Each point of the space represents a label distribution whose expression can be generated. The distributions represent the confidence of neutral, happy, sad, surprised, fear, disgust, and angry (from left to right).

One of the critical features of GANmut is its ability to function as a unified model rather than a set of cross-domain models, paralleling the approach of StarGAN. StarGAN is renowned for its scalability in multi-domain image-to-image translation, enabling modifications of attributes such as hair color, gender, and age within a unified framework. Similarly, GANmut leverages a single, comprehensive model to handle the gamut of human emotions, thus ensuring a streamlined and efficient process for emotion manipulation and generation.

GANmut's ability to maintain the identity of the original image while manipulating the emotional expression is comparable to the identity-preserving qualities of G2GAN. G2GAN employs a dual-GAN architecture to remove and then synthesize expressions while preserving the subject's identity. GANmut, on the other hand, achieves this through a sophisticated learning process that generates a continuous and interpretable conditional space for emotions, ensuring that the synthetic images remain true to the original identities.

The Multitask Cascaded Convolutional Networks (MTCNN) model [20], introduced by Zhang et al., has proven to be a robust and efficient method for joint face detection and alignment in unconstrained cases. This model addresses the challenges posed by various facial poses, illuminations, and occlusions through a deep learning framework that exploits the inherent correlation between detection and alignment tasks. The MTCNN employs a cascaded architecture comprising three stages of deep convolutional networks, which refine the detection and alignment process in a coarse-to-fine manner.

Initially, a shallow CNN generates candidate facial windows, which are then refined and filtered through subsequent, more complex CNN stages. This approach not only enhances accuracy but also maintains real-time performance. Additionally, the MTCNN incorporates an innovative online hard sample mining strategy, which dynamically improves the model's robustness by focusing on difficult samples during training. Extensive evaluations on challenging benchmarks, such as the FDDB and WIDER FACE datasets for face detection and the AFLW dataset for face alignment, demonstrate that MTCNN outperforms state-of-the-art methods, providing superior accuracy and efficiency. This makes it an ideal choice for preprocessing synthetic images in FER applications, ensuring that the synthetic datasets are well-aligned with natural datasets like RAF-DB.

The POSTER (Pyramid crOss-fuSion TransformER) model [21] represents a significant advancement in the field of Facial Expression Recognition (FER) by addressing three critical challenges: inter-class similarity, intra-class discrepancy, and scale sensitivity. Unlike traditional approaches that either use handcrafted features or deep learning methods focusing on a single aspect of these challenges, POSTER integrates a two-stream architecture comprising both image and landmark streams. The model leverages a transformer-based cross-fusion method that allows effective collaboration between facial landmark features and image features. This design ensures that salient facial regions receive proper attention, significantly enhancing the model's ability to differentiate between similar expressions and mitigate discrepancies within the same expression category. Furthermore, the pyramid structure incorporated in POSTER promotes scale invariance, enabling the model to perform consistently across varying image resolutions. Extensive experimental results have demonstrated POSTER's superiority, achieving state-of-the-art accuracy on prominent FER datasets such as RAF-DB (92.05%), FERPlus (91.62%), and AffectNet (67.31% for 7-class and 63.34% for 8-class tasks).

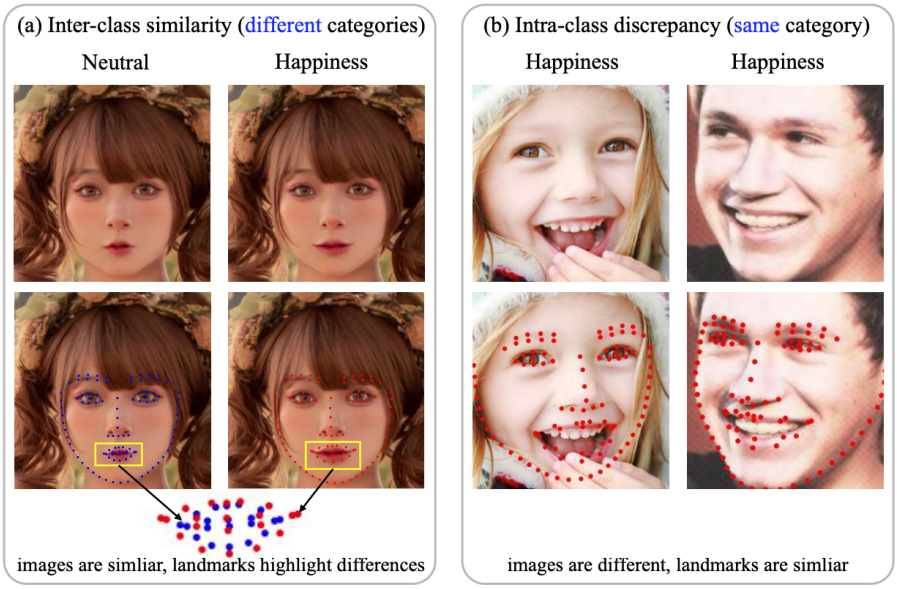

Figure 2.6: Inter-class similarity and intra-class discrepancy. (The facial landmarks are detected by [22].)

The preference for POSTER in this project is driven by its holistic approach to solving key issues in FER and its demonstrated effectiveness in enhancing classification performance. POSTER's ability to combine global image features with localized landmark features through a cross-fusion transformer network provides a robust framework for recognizing subtle differences in facial expressions. This is particularly important in FER, where minor variations in facial regions can significantly alter the perceived emotion. By addressing the limitations of scale sensitivity, POSTER ensures consistent performance across diverse image qualities and resolutions, which is crucial for real-world applications. The model's innovative architecture not only improves accuracy but also offers a comprehensive solution to the challenges that have hindered FER performance in the past. These capabilities make POSTER an ideal choice for training and testing FER models, facilitating a detailed comparative analysis between natural and synthetic data. This analysis aims to validate the efficacy of synthetic data in FER applications, potentially revolutionizing the way facial expression datasets are created and utilized.

3 Developed Approach and System Model

In this section, we present a comprehensive overview of the developed approach and system model for enhancing Facial Expression Recognition (FER) using synthetic data. The methodology encompasses the generation of synthetic datasets through the GANmut model, addressing domain adaptation challenges with advanced preprocessing techniques, and leveraging state-of-the-art model architectures like POSTER and MTCNN for robust FER. This holistic approach aims to mitigate the limitations of existing FER datasets, improve model generalization, and establish the efficacy of synthetic data in training deep learning models for FER applications.

3.1 Generation of Synthetic Data

The generation of synthetic data plays a pivotal role in this project, which aims to address the inherent limitations and biases present in existing facial expression recognition (FER) datasets. The primary purpose of generating synthetic data is to supplement the available natural data, thereby enhancing the diversity and volume of the training datasets used for FER models. This approach is intended to mitigate the challenges posed by insufficient training data, which often leads to overfitting and biased model performance. By integrating synthetic data, we aim to create more robust and generalizable FER models capable of performing accurately across diverse real-world scenarios.

The synthetic data generation process begins with the GANmut model, a GAN-based framework specifically designed for emotion editing on natural images. GANmut learns an expressive and interpretable conditional space of emotions, enabling the generation of synthetic facial expressions that are both realistic and diverse. This process is crucial for creating high-quality synthetic datasets that can effectively complement existing natural datasets like the RAF-DB. Two synthetic datasets, SynFace15F and SynFace15V, were generated to facilitate a comparative analysis. SynFace15F represents the initial version without domain adaptation, while SynFace15V includes advanced techniques like color transfer in the LAB color space and alpha blending to ensure the synthetic images closely match the natural ones in terms of skin tone and other domain-specific features. Additionally, the resolution of the synthetic datasets was matched to the natural dataset's resolution of 100x100 pixels to avoid any external side effects on training performance due to the synthesis process we have implemented.

The primary goal of this comparative analysis is to evaluate the performance of FER models trained on natural data against those trained on synthetic data. By comparing the classification accuracy of models trained with RAF-DB (natural dataset), SynFace15F, SynFace15V, and a hybrid dataset (RafSynFace30V), we aim to determine the viability of using synthetic data as a substitute for natural data in training FER models. The comparative analysis also seeks to identify any significant performance gaps between models trained on different types of data, thereby highlighting the impact of synthetic data generation techniques on model performance.

An essential aspect of this comparative analysis is the need to match the class sample distribution of the synthetic datasets with that of the natural dataset, RAF-DB. Ensuring that both datasets have an identical distribution of emotion classes is crucial for a fair comparison of model performance. Discrepancies in class sample distribution can lead to biased results, as models may perform better in overrepresented classes while underperforming in underrepresented ones. By generating the same number of images for each emotion class as present in the RAF-DB dataset, we ensure that the synthetic datasets are comparable to the natural dataset in terms of class distribution. This approach allows for a more accurate assessment of the models' ability to generalize across different types of data and provides a clearer understanding of the benefits and limitations of using synthetic data in FER.

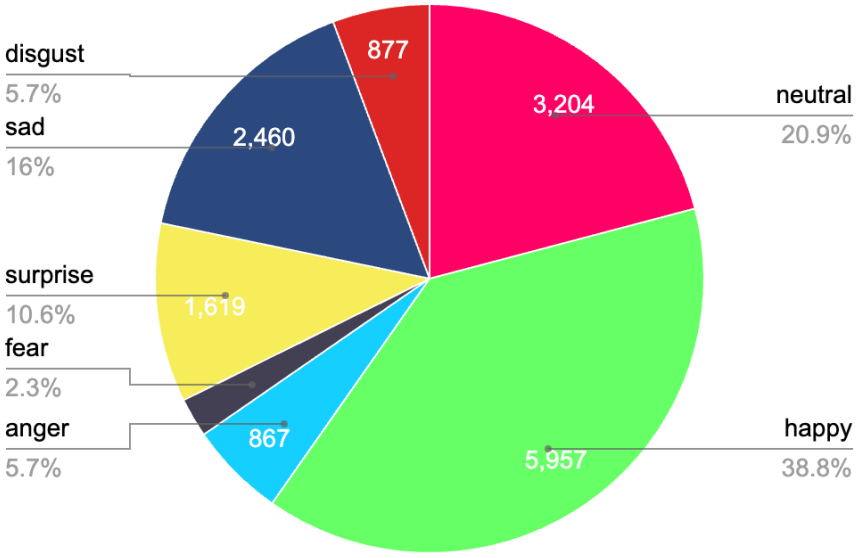

Figure 3.1: Distribution of synthetic images generated for each emotion class.

3.1.1 Conditional Space of Emotions and Variance of Emotion Intensity

To avoid overfitting to specific points within the conditional space or gamut of emotions, the regions from which emotions are synthetically generated were selected from a randomly defined range. This range was determined through a manual inspection process, where emotions generated at various points within the conditional space were visually assessed to ensure a diverse and representative selection. By sampling from a broad range of values, we ensure that the synthetic data covers a wide spectrum of emotional expressions, thereby enhancing the robustness and generalizability of the resulting FER models. This approach mitigates the risk of the models becoming too finely tuned to specific, potentially non-representative points in the emotional spectrum, which could otherwise lead to overfitting and reduced performance in real-world applications.

In the context of controlling emotion intensity and distinguishing between different emotion classes, we adopted the rho-theta representation of the conditional space of emotions. The rho value is particularly effective for modulating the intensity of emotions, allowing for the generation of expressions with varying degrees of strength. Conversely, the theta value serves to differentiate between distinct emotion classes, ensuring clear and discernible boundaries between different types of expressions. By leveraging the rho-theta representation, GANmut facilitates precise and flexible control over the generated emotions, resulting in synthetic data that not only spans a comprehensive range of emotional intensities but also maintains distinctiveness across different emotion classes. This method enhances the quality and utility of the synthetic data, making it a valuable asset for training more accurate and robust FER models.

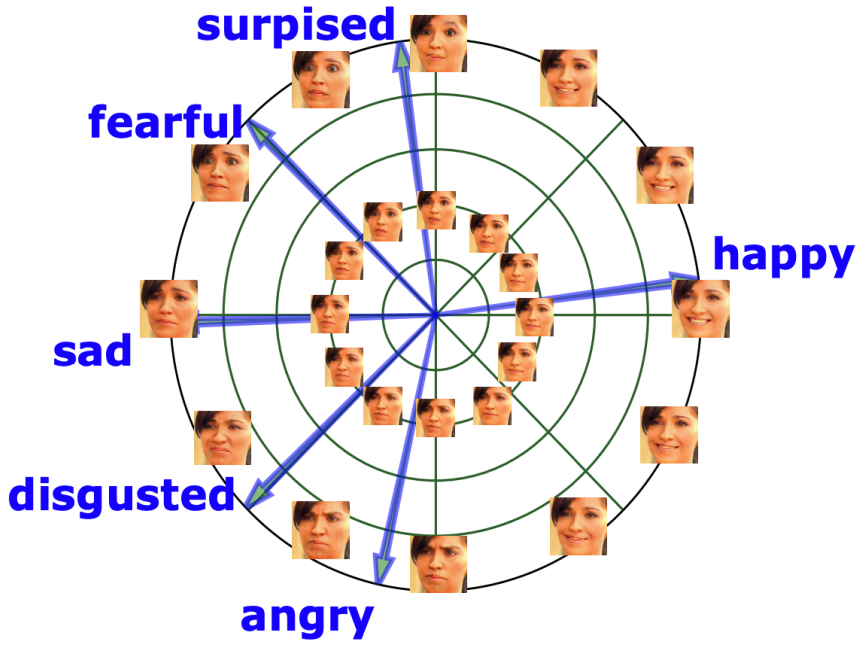

Figure 3.2: Conceptual illustration of a Gamut of emotions in rho-theta coordinate space. We learn to generate diverse/complex emotions merely using the labels of basic categorical emotions (blue arrows).

3.1.2 Color Transfer for Domain Adaptation

After the GANmut model performs emotion adjustment, a notable challenge arises due to slight changes in the skin color of the face. These changes can create a domain adaptation problem when the edited and emotion-adjusted face region is pasted back into the original image, resulting in unnatural visual artifacts. To address this issue, we implemented a color transfer technique in the CIELAB color space, which is specifically designed to handle color matching and adaptation tasks more effectively than other color spaces.

The preference for using the CIELAB color space in our color transfer process stems from its perceptual uniformity properties. Unlike the RGB color space, where changes in color components do not correspond linearly to perceived changes in color, the CIELAB color space is designed to be more aligned with human vision. This means that Euclidean distances in the CIELAB space correspond more closely to perceived color differences, making it an ideal choice for tasks that require precise color matching. By leveraging the CIELAB color space, we ensure that the transferred colors look more natural and consistent to the human eye, thus effectively blending the edited face region with the rest of the image.

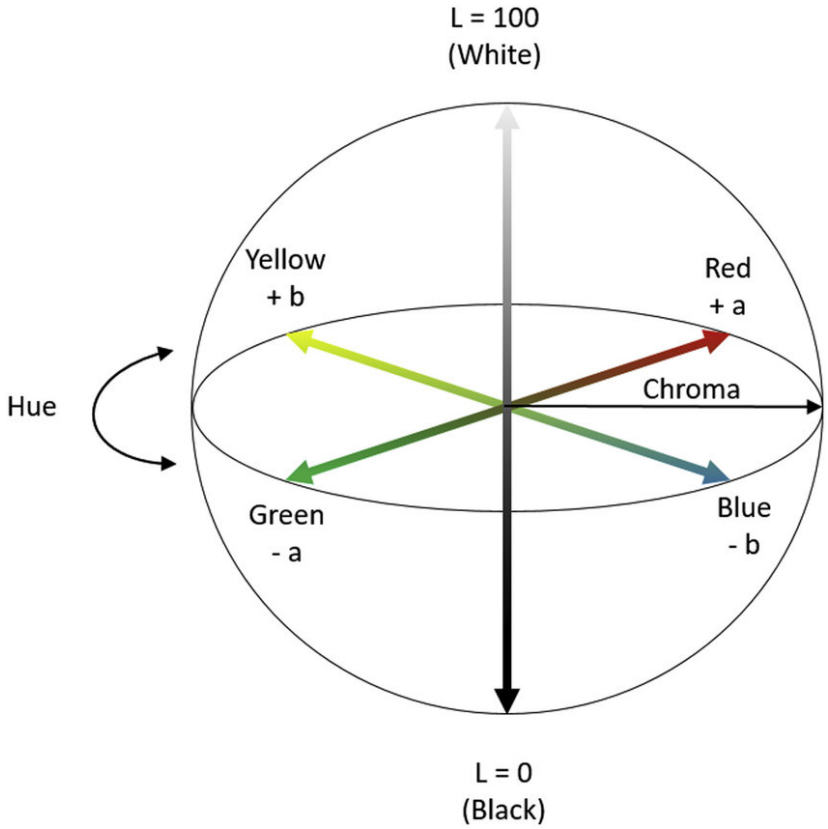

Figure 3.3: The CIELAB color space diagram. [23]

The purpose of the color transfer process is to match the face skin color of the edited region with that of the original image, ensuring a seamless integration. The process involves several steps: first, converting both the source image (edited face region) and the target image (original face region) to the CIELAB color space. This is followed by computing the mean and standard deviation of each color channel in the LAB space for both images. These statistics are then used to adjust the color distribution of the target image to match that of the source image. Specifically, we subtract the mean of the target image, normalize it by its standard deviation, then scale it by the source image's standard deviation and add the source mean. This adjustment aligns the color properties of the target image with those of the source, effectively transferring the color.

The final step involves clipping the adjusted image to ensure that all pixel values fall within the valid range [0, 255] and converting the result back to the RGB color space. This process ensures that the skin color in the edited face region matches the original image's skin color, thereby eliminating the domain adaptation problem and resulting in a more natural and visually coherent image. The use of CIELAB space for this color transfer process significantly improves the visual quality of the synthetic data, making it more suitable for training FER models.

3.1.3 Alpha Blending for Domain Adaptation

The second step to resolve the domain adaptation problem involves ensuring a smooth transition after cropping the detected face for further emotion editing by GANmut and pasting the facial region back into the original image post-adjustment. Crop and paste operations, along with the modifications made by GANmut to achieve the target facial expression, can result in visual discrepancies between the edited region and the natural image. These discrepancies, if not addressed, can lead to distinguishable differences that undermine the realism of the synthetic images. To mitigate this issue, we incorporated an alpha blending step into our pipeline following the color transfer application.

Alpha blending is a technique used to blend two images smoothly, ensuring that the transition between them appears natural. In our context, this process involves creating an alpha mask that gradually transitions from fully opaque to fully transparent at the borders of the cropped region. The mask is applied during the blending of the edited facial region with the original image, ensuring that the edges of the pasted region blend seamlessly with the surrounding pixels. This method helps maintain the natural appearance of the image, eliminating harsh edges or noticeable differences that could arise from the cropping and pasting operations.

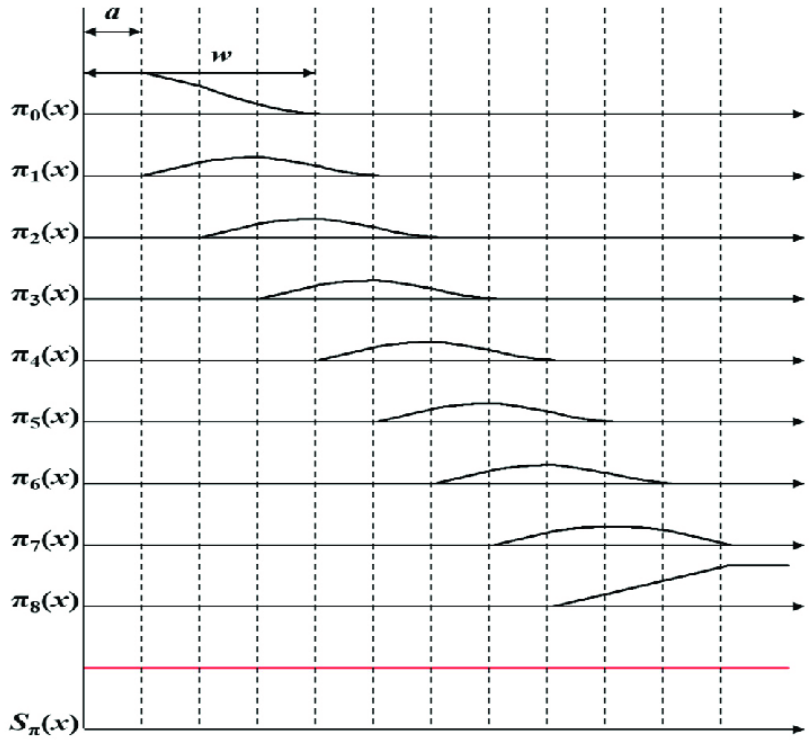

Figure 3.4: Alpha-blended signal in a shifted window function format. [24]

The alpha blending process begins by creating an alpha mask for the blended region. The mask is designed with a defined border width for linear interpolation, ensuring a gradual transition at the edges. The center of the mask is filled with a value of 1.0, indicating full opacity, while the borders are interpolated to transition smoothly to transparency. This mask is then merged into a 3-channel image to match the color channels of the original and edited images. The original region of the image where the face will be pasted is extracted and converted to a floating-point format to facilitate the blending operation. The actual alpha blending is performed by combining the alpha mask with the resized face (post-GANmut adjustment) and the original region, ensuring a seamless transition between the two. The final blended region is then inserted back into the original image, resulting in a smooth and natural integration of the edited facial region.

This alpha blending step is crucial for maintaining the visual coherence of the synthetic images. By smoothing the transitions between the edited and original regions, it ensures that the synthetic data closely resembles natural images, thereby reducing any potential domain adaptation issues. This enhancement, combined with the color transfer step, significantly improves the quality and realism of the synthetic datasets, making them more effective for training robust and accurate FER models.

3.2 Model Architectures

3.2.1 GANmut

The GANmut model represents a significant innovation in the generation of synthetic facial expressions by learning an expressive and interpretable conditional space for a gamut of emotions. Unlike traditional conditional GANs that rely on handcrafted labels, GANmut uses basic categorical emotion labels to jointly learn conditional space and emotion manipulation. This framework leverages a two-dimensional polar coordinate system, where each emotion is represented as a vector originating from a neutral point. The angle (theta) of the vector corresponds to the type of emotion, while the length (rho) indicates the intensity of the emotion. This representation allows GANmut to generate a wide range of complex and compound emotions by sampling points along and between these vectors, providing a nuanced control over both emotion type and intensity.

The architecture of GANmut integrates a generator and discriminator in an adversarial setup, augmented by additional loss functions to enhance the interpretability and quality of the generated images. The generator is tasked with producing realistic facial expressions that align with the desired emotional attributes, while the discriminator evaluates the authenticity of the generated images and their adherence to the specified emotions. Key components of the model include adversarial loss, classification loss, and regression loss, which collectively ensure that the generated images are realistic, correctly classified, and accurately mapped to the intended conditional space. This joint optimization approach not only enhances the model's ability to produce realistic facial expressions but also ensures a smooth transition and meaningful interpolation between different emotional states.

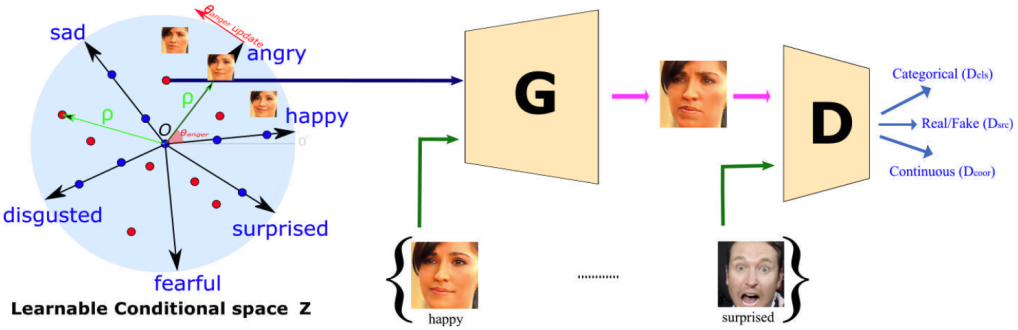

Figure 3.5: Overview of the proposed GANmut model. During the training, one part of the batch is conditioned with codes randomly sampled from Z (red points) and the other (blue points) sampled in the proximity of one of the learnable vectors (representing basic emotion). Only the second part of the batch undergoes the classification loss (for categorical labels) so that red points are free to encode any expression difficult to describe by basic emotions. All points are expected to generate realistic faces.

During the training of GANmut, the data batch is split into two parts. The first part of the batch is conditioned with codes randomly sampled from the conditional space Z. These randomly sampled points (represented as red points) allow the model to explore a wide range of possible expressions, including complex and nuanced emotions that may not be easily classified into basic emotion categories. This strategy enables the model to learn a diverse and rich set of facial expressions that go beyond the predefined basic emotions.

The second part of the batch is conditioned with codes sampled in the proximity of one of the learnable vectors, which represent the basic emotions (represented as blue points). These points are specifically used to train the model with basic categorical emotion labels. The classification loss is applied only to this part of the batch, ensuring that the generated images can be correctly classified according to the basic emotion categories. This focused training helps the model learn to generate clear and distinguishable expressions corresponding to the basic emotions.

By using this dual sampling strategy, GANmut ensures that the entire conditional space is effectively covered. The red points facilitate the learning of complex and intermediate emotions, while the blue points ensure accurate generation of basic emotions. The overall objective is to generate realistic facial expressions for all points in the conditional space, thereby creating a versatile and robust model capable of producing a wide gamut of human emotions.

The preference for the GANmut model in this project is driven by its comprehensive capabilities and innovative features. Firstly, GANmut's ability to learn from basic categorical emotion labels while generating a wide range of complex and compound emotions makes it particularly effective for FER applications. This model does not rely on handcrafted labels, which often come with their own set of limitations and biases. Instead, GANmut learns an interpretable conditional space that can produce diverse emotional expressions, making it a robust tool for generating synthetic facial data.

Secondly, GANmut's unified approach to emotion synthesis aligns well with the project's goals of creating high-quality synthetic datasets. The model's efficiency in generating realistic and diverse expressions without the need for multiple cross-domain models simplifies the data generation process. This efficiency is crucial for developing large-scale synthetic datasets like SynFace15F and SynFace15V, which are integral to this project's methodology.

Additionally, GANmut's photorealistic image generation capabilities ensure that the synthetic data produced is of the highest quality. This quality is essential for training FER models, as the effectiveness of these models is heavily dependent on the quality and diversity of the training data. By generating images that are both realistic and diverse, GANmut enhances the potential for FER models to generalize well across different scenarios and subjects.

Furthermore, GANmut's approach to preserving the identity of the original image while manipulating the emotional expression is particularly valuable for maintaining the integrity of the synthetic data. This capability ensures that the synthetic images are not only realistic but also retain the unique characteristics of the original subjects, making them more useful for training robust FER models.

3.2.2 MTCNN

The Multitask Cascaded Convolutional Networks (MTCNN) model is designed to tackle the challenges of face detection and alignment in unconstrained environments, such as varying poses, illumination conditions, and occlusions. MTCNN adopts a cascaded architecture consisting of three carefully designed stages of deep convolutional networks, each stage refining the detection and alignment process in a coarse-to-fine manner. The first stage, known as the Proposal Network (P-Net), quickly generates candidate facial windows and their bounding box regression vectors. These candidates are then refined and calibrated through non-maximum suppression (NMS) to eliminate highly overlapped candidates. The second stage, the Refine Network (R-Net), further filters out false candidates, performs bounding box regression and conducts another round of NMS. Finally, the Output Network (O-Net) refines the detection results and outputs the positions of five facial landmarks, ensuring precise alignment.

The architecture of MTCNN leverages multitask learning to enhance the performance of both face detection and alignment tasks simultaneously. Each network within the cascade is trained using a combination of loss functions tailored to different objectives: face classification, bounding box regression, and facial landmark localization. For face classification, the network uses a cross-entropy loss, while bounding box regression and landmark localization are optimized using Euclidean loss functions. An innovative aspect of MTCNN is its online hard sample mining strategy, which adaptively focuses on hard-to-classify samples during training, thus improving the robustness and accuracy of the model. This multitasking approach, combined with the cascaded architecture, enables MTCNN to achieve superior accuracy and real-time performance on challenging benchmarks such as FDDB and WIDER FACE for face detection and AFLW for face alignment.

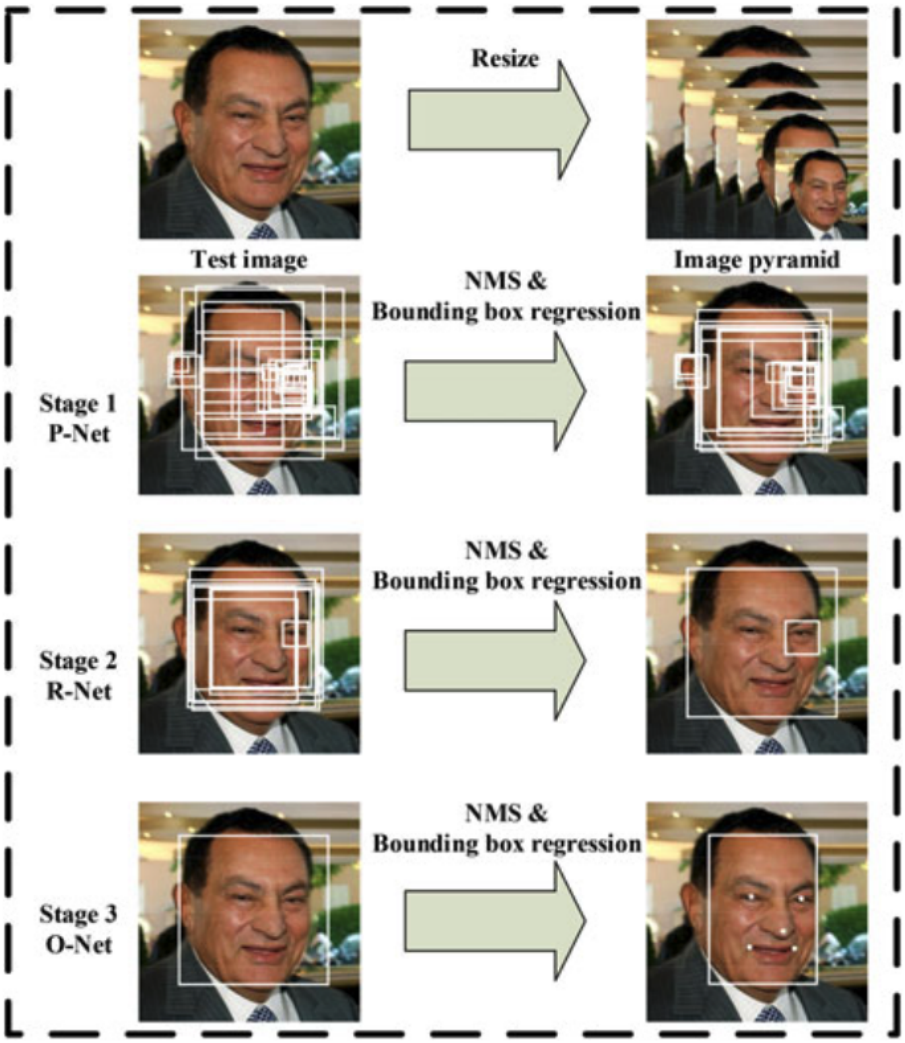

Figure 3.6: Pipeline of the cascaded framework of MTCNN that includes three-stage multitask deep convolutional networks.

3.2.3 POSTER

The POSTER (Pyramid crOss-fuSion TransformER) model is a sophisticated architecture designed to address three critical challenges in facial expression recognition (FER): inter-class similarity, intra-class discrepancy, and scale sensitivity. The model employs a two-stream architecture that integrates both image and landmark streams, leveraging a transformer-based cross-fusion mechanism to enhance feature representation. In this design, facial landmarks pinpoint salient facial regions and guide the model's attention, effectively addressing the issue of inter-class similarity by focusing on the subtle differences that distinguish various expressions. Simultaneously, the global context provided by image features helps to mitigate intra-class discrepancies, ensuring that variations within the same expression category, such as differences in skin tone, gender, and age, are adequately captured.

POSTER's architecture consists of a pyramid structure to promote scale invariance, a critical factor given the varying resolutions and quality of images in FER datasets. The pyramid structure allows the model to capture feature maps at multiple scales, enhancing its robustness to different image sizes and ensuring consistent performance across diverse scenarios. The cross-fusion transformer blocks are central to POSTER's design, where image and landmark features are processed in parallel streams. By swapping key matrices between these streams, POSTER enables a mutual enhancement of feature representations, allowing the model to integrate both global and local information effectively. This design ensures that salient regions highlighted by landmarks are enriched with global context from the image features, improving the model's overall discriminative power.

The training of POSTER involves fine-tuning the image backbone while keeping the landmark detector frozen to maintain accurate landmark outputs. The cross-fusion mechanism is implemented within the multi-head self-attention layers of the transformer, where the attention mechanism facilitates the integration of complementary features from both streams. This approach not only enhances the model's ability to handle the subtle nuances of facial expressions but also ensures that the learned representations are robust and generalizable. Extensive experiments have demonstrated that POSTER achieves state-of-the-art results on several FER benchmarks, including RAF-DB, FERPlus, and AffectNet, showcasing its effectiveness in tackling the complex challenges of FER.

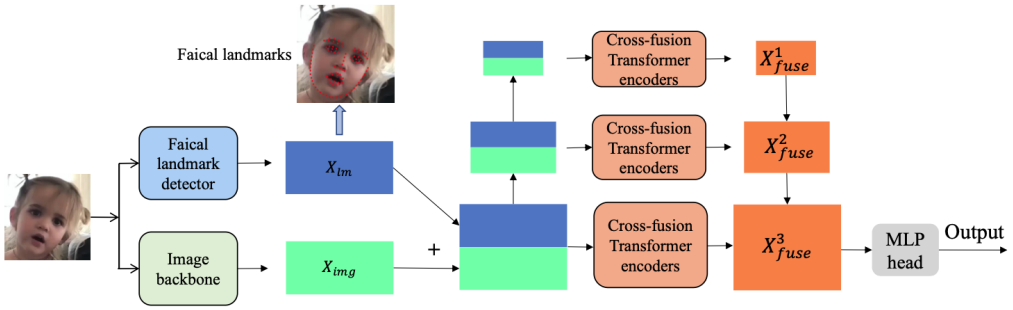

Figure 3.7: The architectures of proposed POSTER for Facial Expression Recognition (FER). A facial landmark detector (MobileFaceNet [7]) is applied to obtain landmark features Xlm. The image backbone (IR50 [10]) is used to extract image features Ximg. "+" denotes patch-wise concatenation operation.

The choice of POSTER for our comparative analysis between natural and synthetic datasets is driven by its superior performance and comprehensive approach to FER. POSTER's ability to address inter-class similarity, intra-class discrepancy, and scale sensitivity makes it an ideal candidate for evaluating the effectiveness of synthetic data. By training POSTER on both natural datasets (like RAF-DB) and synthetic datasets (such as SynFace15F and SynFace15V), we can rigorously assess the impact of synthetic data on model performance. The state-of-the-art accuracy achieved by POSTER on diverse benchmarks ensures that any observed differences in performance are likely attributable to the quality of the training data rather than the limitations of the model itself.

Moreover, POSTER's advanced feature integration through its cross-fusion transformer blocks provides a robust framework for analyzing the nuances of facial expressions, making it particularly well-suited for this study. The model's ability to leverage both image and landmark features ensures that it can effectively utilize the rich, diverse data generated by synthetic processes.

4 Experimentation Environment and Experiment Design

The experimentation environment for this project is designed to rigorously evaluate the effectiveness of synthetic data in enhancing Facial Expression Recognition (FER) model performance. The core objective is to compare the classification accuracies of FER models trained on natural datasets versus those trained on synthetic datasets. The detailed experimentation setup and design are structured to ensure a comprehensive analysis, leveraging advanced techniques and robust model architectures to achieve reliable and insightful results.

4.0.1 Data Generation and Preprocessing

The initial step in our experimental pipeline involves the generation of synthetic facial expression data using the GANmut model. The synthetic data is generated from the FFHQ Flickr-Faces-HQ Dataset, producing the SynFace15F dataset. The GANmut model operates within an expressive and interpretable conditional space, where it generates diverse facial expressions by manipulating key parameters. This generation process ensures that the synthetic dataset encompasses a wide range of emotional expressions, providing a robust foundation for subsequent model training.

Addressing the domain adaptation problem is critical to ensure the synthetic data's compatibility with natural datasets. After GANmut adjusts emotions, a two-step domain adaptation process is implemented. The first step involves a color transfer in the CIELAB color space. This technique is chosen for its perceptual uniformity, allowing for precise color matching that aligns the skin tones of the edited facial regions with the original images. The second step, alpha blending, ensures a smooth transition between the edited and original facial regions. This technique mitigates the visual discrepancies that can arise from the crop-and-paste operations inherent in the emotion editing process. The result is the SynFace15V dataset, which exhibits enhanced visual coherence and realism.

4.0.2 Dataset Preparation

To facilitate a fair comparison, it is crucial to match the class sample distribution and resolution of the synthetic datasets with the natural dataset. The RAF-DB dataset, a widely used natural dataset in FER research, serves as our benchmark for this purpose. The resolution of the synthetic datasets is adjusted to match the 100x100 pixel resolution of RAF-DB, ensuring uniformity across all datasets. Additionally, the class sample distribution is aligned to mirror that of RAF-DB, with equal representation for each of the seven emotion classes (neutral, happy, angry, fear, surprise, sad, disgust). This alignment is essential for mitigating potential biases and ensuring a fair and accurate comparative analysis.

4.0.3 Datasets

Dataset	Purpose	Type	Size
FFHQ	Generation	Natural	70.000
RAF-DB	Training & Testing	Natural	15.339
SynFace15F	Training & Testing	Synthetic	15.339
SynFace15V	Training & Testing	Synthetic	15.339
RafSynFace30V	Training & Testing	Hybrid	30.678
AffectNet-Val	Testing	Natural	3.500

Figure 4.1: Table of the datasets utilized or synthesized during the whole process of this research.

The experimentation environment for this project utilizes a diverse set of datasets to ensure a comprehensive evaluation of the impact of synthetic data on facial expression recognition (FER) model performance. These datasets include both natural and synthetic images with varying purposes and sizes, as summarized in the provided table.

The FFHQ (Flickr-Faces-HQ) dataset, comprising 70,000 high-quality natural images, serves as the foundational source for generating synthetic data using the GANmut model. This dataset provides a rich and diverse set of facial images, which are crucial for training GANmut to produce realistic synthetic facial expressions. The RAF-DB dataset, containing 15,339 natural images, is used for both training and testing FER models. It serves as the benchmark natural dataset against which the performance of models trained on synthetic data is compared.

The synthetic datasets generated include SynFace15F and SynFace15V, each containing 15,339 images. SynFace15F represents the initial version generated directly by GANmut, while SynFace15V includes additional steps of color transfer in the CIELAB color space and alpha blending to address domain adaptation issues. These synthetic datasets are designed to match the class distribution and resolution of the RAF-DB dataset to facilitate a fair comparison. Additionally, a hybrid dataset, RafSynFace30V, combines RAF-DB and SynFace15V, resulting in a dataset of 30,678 images. This hybrid dataset aims to explore the potential benefits of combining natural and synthetic data for training FER models. Finally, the AffectNet validation dataset, consisting of 3,500 natural images (500 images per emotion class), is used exclusively for testing. This independent dataset provides a robust benchmark for evaluating the generalization capabilities of the trained models, free from any biases inherent in the training datasets.

4.0.4 Training and Model Architecture

The POSTER (Pyramid crOss-fuSion TransformER) model is employed for training on the prepared datasets. POSTER is selected for its state-of-the-art performance and its robust approach to handling the nuances of facial expression recognition. The model integrates both image and landmark streams, leveraging a cross-fusion transformer mech-

anism to enhance feature representation. This architecture allows POSTER to effectively address inter-class similarity, intra-class discrepancy, and scale sensitivity—key challenges in FER.

Four separate models are trained using the following datasets: RAF-DB (natural), SynFace15F, SynFace15V, and RafSynFace30V (a hybrid dataset combining RAF-DB and SynFace15V). The training process involves fine-tuning the image backbone of POSTER while maintaining the accuracy of facial landmark detections. This ensures that the models are optimized for both global and local feature representations, enhancing their ability to recognize subtle variations in facial expressions.

During the training phase, several parameters are set to ensure optimal performance. The data transformations include random horizontal flips, resizing to 224x224 pixels, normalization, and random erasing with a scale of 0.02 to 0.1. The batch size is set to 64 for training and 32 for validation. The initial learning rate is 0.000015, with an exponential learning rate scheduler applied to adjust the learning rate during training. The training process spans 50 epochs, utilizing the Adam optimizer with a momentum of 0.9. The models are trained on GPUs specified by the 'CUDA_VISIBLE_DEVICES' environment variable, ensuring efficient use of computational resources.

4.0.5 Comparative Analysis and Testing

The final phase of the experiment involves a comprehensive comparative analysis of the trained models' classification accuracies. Each model is evaluated on five test datasets: RAF-DB, SynFace15F, SynFace15V, RafSynFace30V, and an independent validation dataset from AffectNet. The AffectNet validation dataset comprises 500 images per emotion class, providing a robust benchmark for assessing the models' generalization capabilities. This independent dataset is critical for evaluating the models' performance in unbiased scenarios, free from any inherent biases present in the training datasets.

4.0.6 Evaluation Metrics

The evaluation metrics focus on classification accuracy, with detailed analysis performed to identify any significant performance gaps between models trained on different types of data. This analysis provides insights into the effectiveness of synthetic data in enhancing FER model performance and highlights the potential benefits and limitations of integrating synthetic data into FER training pipelines.

For testing, the data transformations include resizing to 224x224 pixels and normalization, ensuring consistency with the training phase. The batch size for testing is set to 32. Pre-trained weights are loaded into the models to maintain consistency in evaluation. The testing process calculates the classification accuracy, with additional metrics like F1-score and confusion matrix to provide a detailed assessment of model performance. The testing environment leverages the same GPU resources as the training phase, ensuring that the evaluations are conducted under consistent computational conditions.

The comprehensive analysis aims to identify any significant performance gaps between models trained on different datasets, thereby highlighting the impact of synthetic data on FER model performance.

4.0.7 Overall Pipeline of the Experiment

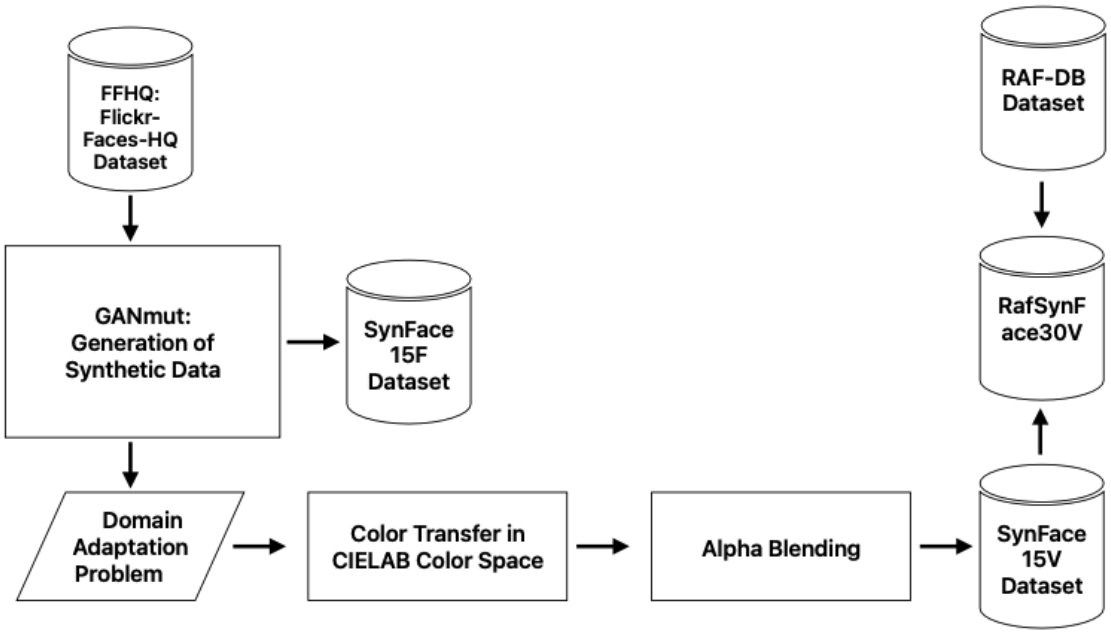

Figure 4.2: Overview of the proposed experiment with cascaded data generation pipeline that illustrates the preparation of SynFace15F, SynFace15V and RafSynFace30V.

The provided diagram illustrates the overall project pipeline, detailing the sequential processes involved in data generation, domain adaptation, and dataset preparation. Starting with the GANmut model, synthetic images are generated from the FFHQ dataset, followed by domain adaptation through color transfer and alpha blending. The resulting SynFace15V dataset, along with SynFace15F and RAF-DB, is used to create the hybrid RafSynFace30V dataset. These datasets are then utilized to train the POSTER model, with the trained models evaluated on multiple test datasets to perform a rigorous comparative analysis of their classification accuracies.

In conclusion, the experimentation environment and design are meticulously crafted to ensure a robust evaluation of synthetic data's role in FER. Through advanced model architectures, precise data preprocessing, and comprehensive comparative analysis, this project aims to provide valuable insights into the potential of synthetic data to enhance facial expression recognition systems.

5 Comparative Evaluation and Discussion

5.1 Evaluation of Domain Adaptation Improvements of GANmut

POSTER MODEL Training/Test Datasets	SynFace15F (3065)	SynFace15V (3065)	RafSynFace30V (6133)	AffectNet-Val (3500)
SynFace15F	**0.9990**	0.7811	0.7390	0.4931
SynFace15V	0.9896	**0.9856**	**0.8342**	**0.4949**

Figure 5.1: Classification accuracy table for models trained with SynFace15F and SynFace15V datasets. The table shows the accuracy of these models when tested on four different datasets: SynFace15F, SynFace15V, RafSynFace30V, and AffectNet-Val. The results highlight the impact of domain adaptation improvements in the SynFace15V dataset on model performance across various test datasets, demonstrating enhanced generalization and robustness.

The results presented in the table provide a comprehensive evaluation of the impact of domain adaptation improvements in the GANmut model on the classification performance of Facial Expression Recognition (FER) models. The primary focus is to compare the performance of models trained on the SynFace15F dataset, which lacks domain adaptation improvements, against those trained on the SynFace15V dataset, which incorporates two-step domain adaptation and varied emotion intensity generation.

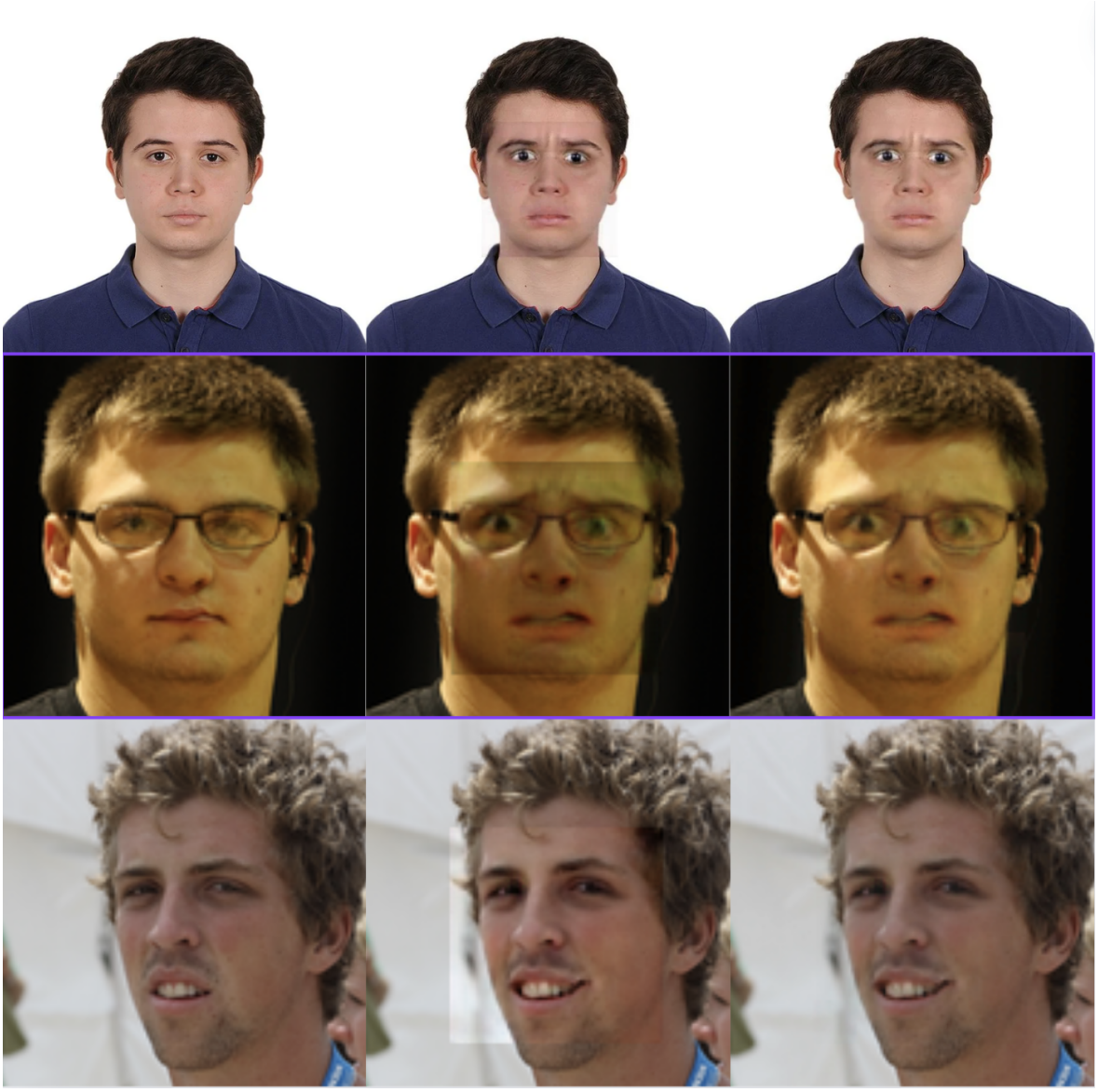

Figure 5.2: Comparison of synthetic images before and after domain adaptation. The leftmost column shows the original images, the middle column displays synthetic images without domain adaptation, and the rightmost column presents synthetic images with domain adaptation using color transfer in CIELAB color space and alpha blending.

The model trained on the SynFace15F dataset shows clear signs of overfitting to specific emotion intensity values within the conditional space. This is evident from its high classification accuracy of 0.9990 on the SynFace15F test set. However, this model's performance drops significantly when evaluated on other datasets. For instance, the accuracy drops to 0.7811 on the SynFace15V test set, 0.7390 on the RafSynFace30V test set, and 0.4931 on the AffectNet-Val test set. These results indicate that while the model performs exceptionally well on data similar to its training set, it struggles with generalization across different datasets, highlighting the limitations of training with fixed emotion intensity values.

In contrast, the model trained on the SynFace15V dataset, which includes domain adaptation improvements through color transfer in the CIELAB color space and alpha blending, as well as varied emotion intensity values, demonstrates superior performance across all test sets. The accuracy of this model on the SynFace15V test set is 0.9856, only slightly lower than the overfitted model on its corresponding test set, but significantly higher than the SynFace15F-trained model on the same dataset. Moreover, the SynFace15V-trained model achieves an accuracy of 0.8342 on the RafSynFace30V test set, showcasing a substantial performance gain over the SynFace15F-trained model. This indicates that the domain adaptation techniques effectively enhance the model's ability to generalize to hybrid datasets that combine natural and synthetic images.

The performance of the SynFace15V-trained model on the AffectNet-Val test set is particularly noteworthy. With an accuracy of 0.4949, it slightly outperforms the SynFace15F-trained model's accuracy of 0.4931. While the performance gap is modest, the results highlight the improvements achieved through domain adaptation. The AffectNet-Val dataset, consisting of 500 images per emotion class, serves as an independent benchmark, free from biases inherent in the training datasets. The better performance of the SynFace15V-trained model on this independent dataset underscores the effectiveness of the domain adaptation improvements in enhancing the model's generalization capabilities.

The most significant performance gap in favor of the SynFace15V-trained model is observed on the RafSynFace30V test set, a hybrid dataset combining RAF-DB and SynFace15V. The accuracy of 0.8342 compared to 0.7390 for the SynFace15F-trained model highlights the substantial benefits of incorporating domain adaptation techniques and varied emotion intensity generation. This improvement can be attributed to the more realistic and diverse synthetic images produced by the GANmut model with domain adaptation, which better complements the natural images in the hybrid dataset, leading to improved model performance.

In summary, the evaluation of the domain adaptation improvements in GANmut demonstrates clear benefits in terms of model generalization and performance. The two-step domain adaptation process, involving color transfer in the CIELAB color space and alpha blending, along with the generation of varied emotion intensity values, significantly enhances the quality and realism of synthetic images. These improvements enable the trained models to perform more effectively across diverse datasets, addressing the limitations of overfitting and ensuring robust FER performance.

5.2 Comparative Analysis of Natural Dataset, Synthetic Dataset, and Hybrid Dataset

POSTER MODEL Training/Test Datasets	RAF-DB (3068)	SynFace15V (3065)	RafSynFace30V (6133)	AffectNet-Val (3500)
RAF-DB	**0.9182**	0.5798	0.7491	0.4971
SynFace15V	0.6829	**0.9856**	0.8342	0.4949
RafSynFace30V	0.9087	0.9765	**0.9426**	**0.5277**

Figure 5.3: Comparison table of the testing performances of three models trained on RAF-DB, SynFace15V, and RafSynFace30V datasets. The table shows classification accuracies on four test datasets: RAF-DB, SynFace15V, RafSynFace30V, and the benchmark AffectNet-Val dataset. The results highlight the models' performance in familiar versus independent datasets, demonstrating the effectiveness of synthetic and hybrid datasets in improving facial expression recognition.

The comparative analysis of FER models trained on natural, synthetic, and hybrid datasets reveals several critical insights into the performance and potential of synthetic data in FER applications. The results table demonstrates that models trained on RAF-DB, SynFace15V, and RafSynFace30V datasets achieve the highest classification accuracies when tested on their respective datasets. This outcome is expected due to the inherent dataset biases and pattern similarities within the same datasets, which provide the models with a familiar and consistent context for classification. Consequently, models tend to perform better on the data they were trained on, showcasing high accuracy scores.

However, the true measure of a model's robustness and generalizability lies in its performance on an independent benchmark dataset. The AffectNet-Val dataset serves this purpose, providing a set of images with no inherent biases or patterns similar to those in the training datasets. The analysis of model performance on AffectNet-Val offers a clearer benchmark for evaluating the impact of natural, synthetic, and hybrid data on FER model accuracy. The most significant finding from this analysis is that incorporating synthetic data into natural datasets significantly enhances classification accuracy. The model trained on the hybrid dataset RafSynFace30V achieves an accuracy of 0.5277 on AffectNet-Val, compared to 0.4971 for the model trained on RAF-DB alone. This improvement underscores the effectiveness of using synthetic data to augment natural datasets, enhancing the overall performance of FER models.

The advantage of using a hybrid dataset is further emphasized by the practical considerations of data collection and annotation. Generating synthetic data is significantly less resource-intensive than collecting and manually annotating natural data. By leveraging synthetic data, researchers can quickly and efficiently expand their training datasets, improving model performance without the substantial time and cost associated with traditional data collection methods. This research highlights the potential for synthetic data to complement natural datasets, offering a viable solution for scaling FER models.

Another key finding from this analysis is the potential of synthetic datasets to serve as substitutes for natural datasets in training FER models. The model trained on SynFace15V, a purely synthetic dataset, achieves an accuracy of 0.4949 on AffectNet-Val, which is remarkably close to the 0.4971 accuracy of the model trained on RAF-DB. This close performance indicates that synthetic datasets, when generated and processed with high-quality techniques, can potentially match or even surpass the effectiveness of natural datasets. As generative models and preprocessing techniques continue to improve, the performance gap between synthetic and natural datasets is likely to diminish further, making synthetic data a compelling alternative for training FER models.

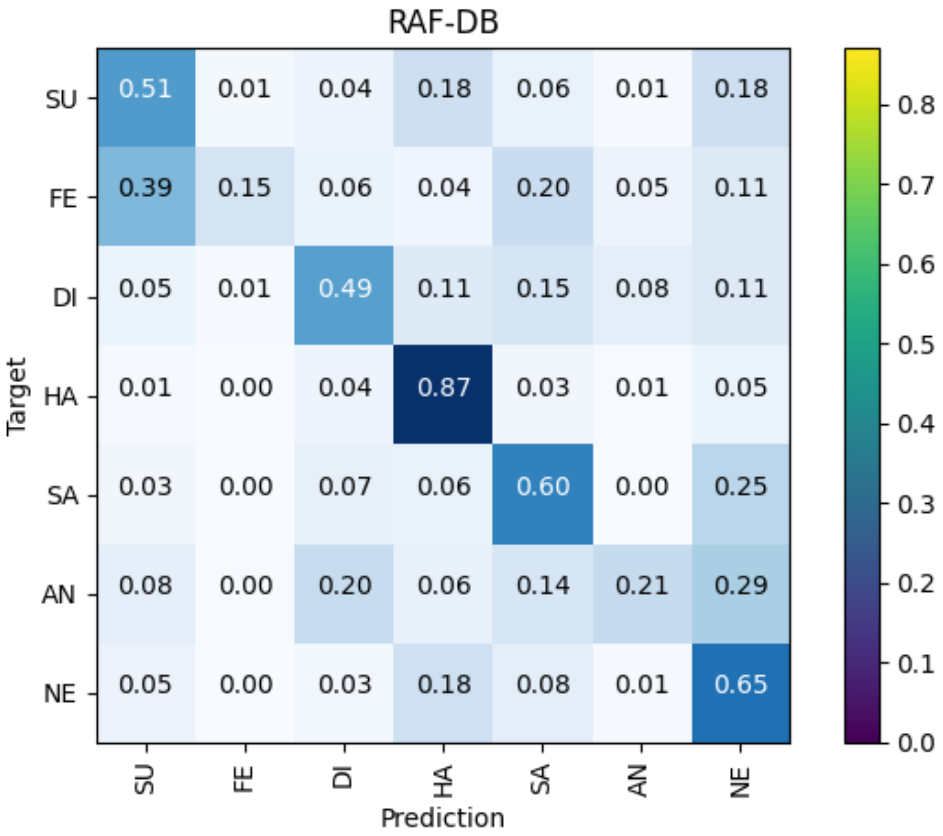

Figure 5.4: Confusion matrices illustrating the classification performance of model trained on RAF-DB dataset when tested on the AffectNet-Val dataset. The matrix shows the distribution of true labels (rows) versus predicted labels (columns) for seven emotion classes: surprise (SU), fear (FE), disgust (DI), happiness (HA), sadness (SA), anger (AN), and neutral (NE). Higher values along the diagonal indicate better classification accuracy for the respective emotion class.

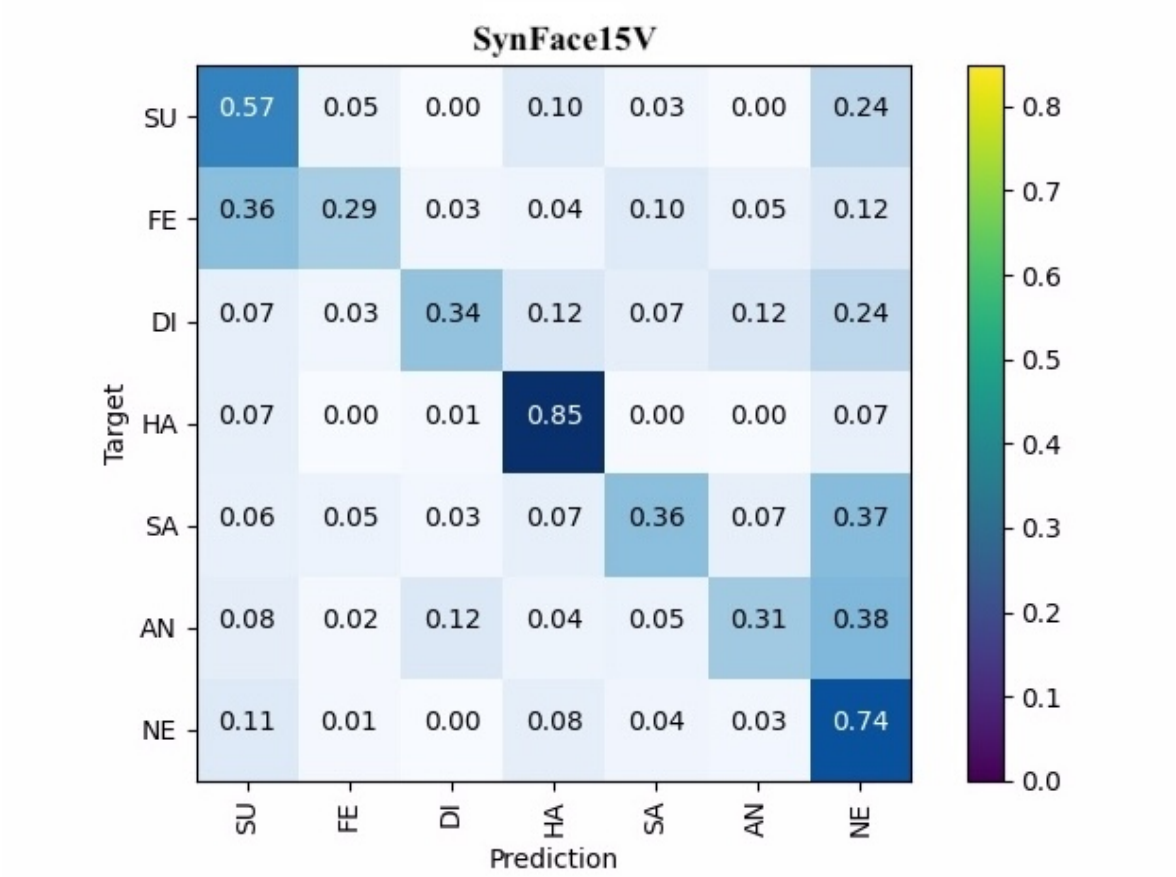

Figure 5.5: Confusion matrices illustrating the classification performance of model trained on SynFace15V dataset when tested on the AffectNet-Val dataset. The matrix shows the distribution of true labels (rows) versus predicted labels (columns) for seven emotion classes: surprise (SU), fear (FE), disgust (DI), happiness (HA), sadness (SA), anger (AN), and neutral (NE). Higher values along the diagonal indicate better classification accuracy for the respective emotion class.

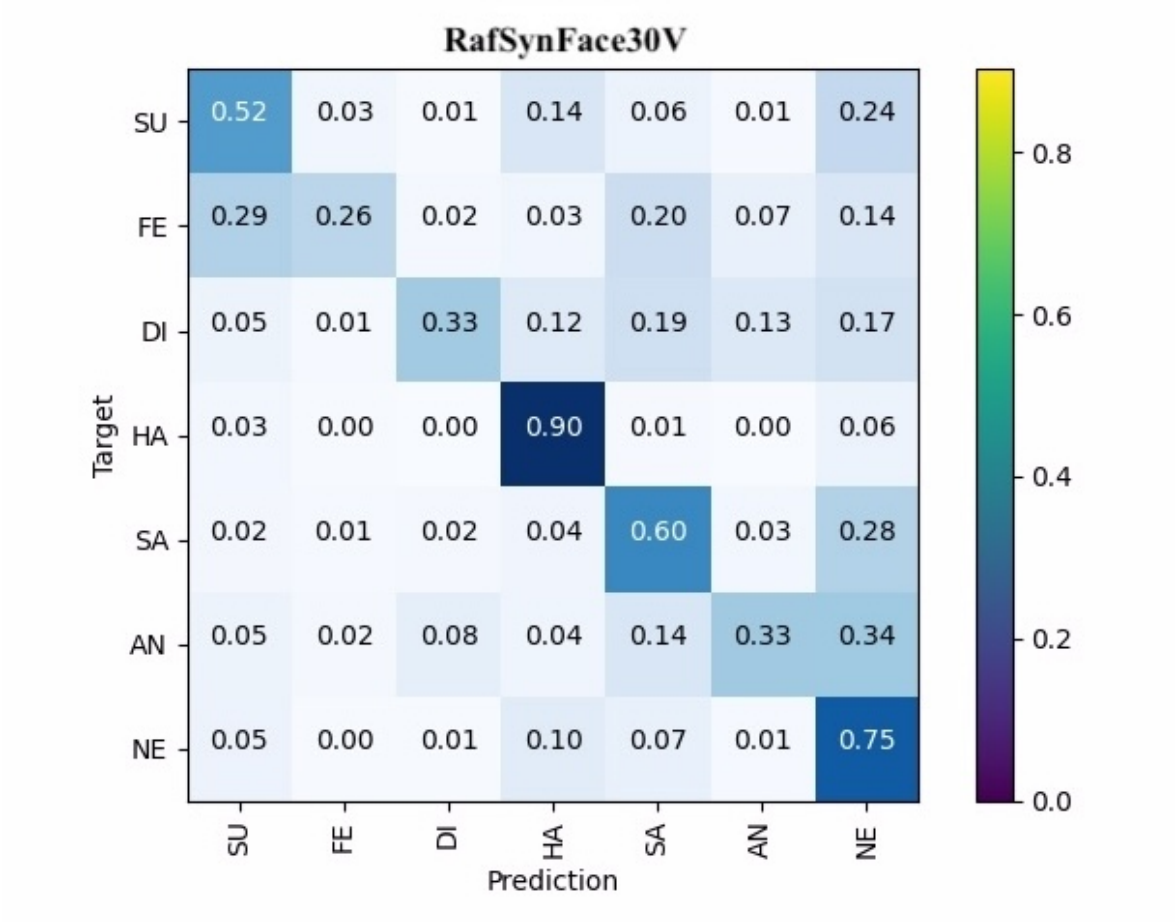

Figure 5.6: Confusion matrices illustrating the classification performance of model trained on RafSynFace30V dataset when tested on the AffectNet-Val dataset. The matrix shows the distribution of true labels (rows) versus predicted labels (columns) for seven emotion classes: surprise (SU), fear (FE), disgust (DI), happiness (HA), sadness (SA), anger (AN), and neutral (NE). Higher values along the diagonal indicate better classification accuracy for the respective emotion class.

The confusion matrices provide a detailed breakdown of the classification performance of models trained on RAF-DB, SynFace15V, and RafSynFace30V when tested on the AffectNet-Val dataset. These matrices reveal the distribution of correct and incorrect predictions across different emotion classes, offering deeper insights into model performance.

For the model trained on RAF-DB, the confusion matrix shows a high degree of confusion between certain emotion classes, such as 'fear' and 'disgust,' and 'anger' and 'sadness.' This indicates that the model struggles to distinguish between these emotions when applied to an independent dataset, likely due to the limitations and biases present in the natural training data.

In contrast, the confusion matrix for the model trained on SynFace15V displays a more balanced distribution of correct predictions across emotion classes. The domain adaptation techniques and varied emotion intensity values used in generating SynFace15V likely contribute to this improved performance, allowing the model to generalize better across diverse expressions.

The confusion matrix for the model trained on RafSynFace30V, the hybrid dataset, shows the highest overall accuracy and the most balanced performance across all emotion classes. The combination of natural and synthetic data in this dataset provides a rich and diverse training set that enables the model to better capture the nuances of facial expressions, resulting in superior generalization capabilities.

In conclusion, the comparative analysis underscores the significant benefits of incorporating synthetic data into FER model training. The hybrid dataset approach offers a practical and effective solution for enhancing model performance, while high-quality synthetic datasets show great potential as standalone training resources. The detailed evaluation and confusion matrices further highlight the improvements in classification accuracy and generalization achieved through these methodologies, paving the way for more robust and scalable FER systems.

6 Conclusion and Future Work

The primary objective of this project was to enhance the field of Facial Expression Recognition (FER) by leveraging synthetic data generated using Generative Adversarial Networks (GANs). Our approach involved the generation of high-quality synthetic facial expression data to address the limitations of existing FER datasets, which are often constrained by their limited size and inherent biases. By integrating synthetic data with natural datasets, we aimed to improve the performance and generalizability of FER models.

Through rigorous experimentation, we demonstrated that the inclusion of synthetic data significantly enhances the robustness and accuracy of FER models. Our comparative analysis showed that models trained on hybrid datasets, combining both natural and synthetic data, outperformed those trained solely on natural data. Specifically, the hybrid dataset RafSynFace30V, which combines the RAF-DB dataset with the SynFace15V synthetic dataset, achieved superior classification accuracy across multiple test datasets, including an independent benchmark dataset, AffectNet-Val. This finding underscores the potential of synthetic data to complement and enhance natural datasets, ultimately leading to better-performing FER models.

The results of our experiments also highlighted the importance of addressing domain adaptation issues in synthetic data generation. By employing techniques such as color transfer in the CIELAB color space and alpha blending, we were able to mitigate domain discrepancies and improve the realism of synthetic images. The refined synthetic dataset, SynFace15V, demonstrated significant performance improvements over the initial version, SynFace15F, particularly when evaluated on independent and hybrid datasets. This reinforces the notion that careful preprocessing and adaptation techniques are crucial for the effective utilization of synthetic data in training deep learning models.

In addition to demonstrating the efficacy of synthetic data, our research contributes to the broader field of FER by providing a comprehensive framework for generating and integrating synthetic facial expression data. The methodologies and insights derived from this project can serve as a foundation for future research and applications in FER and other related domains.

Looking ahead, the next phase of this research will involve scaling the experiments to larger datasets. By utilizing more extensive natural and synthetic datasets, we aim to further validate our findings and reduce the potential impact of internal dataset biases on model performance. Larger datasets will provide a more robust evaluation of the generalization capabilities of FER models trained on synthetic data, offering deeper insights into the scalability and practical applicability of our approach.

Moreover, future work will explore the development of more sophisticated generative models and preprocessing techniques to enhance the quality and diversity of synthetic data. By continuously refining the synthetic data generation process, we aim to push the boundaries of what is achievable with FER models, ultimately contributing to the advancement of artificial intelligence and computer vision technologies.

In conclusion, this project has successfully demonstrated the potential of synthetic data to improve FER models, providing a valuable contribution to the field. The insights gained from this research pave the way for future studies that will further explore and expand the use of synthetic data in various applications, enhancing the accuracy and robustness of machine learning models in real-world scenarios.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[2] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1283–1292, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:21011865

[3] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *CoRR*, vol. abs/1711.09020, 2017. [Online]. Available: http://arxiv.org/abs/1711.09020

[4] S. d'Apolito, D. P. Paudel, Z. Huang, A. Romero, and L. V. Gool, "Ganmut: Learning interpretable conditional space for gamut of emotions," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 568–577.

[5] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 09 2017.

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," 06 2018, pp. 8789–8797.

[7] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[8] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *CoRR*, vol. abs/1708.03985, 2017. [Online]. Available: http://arxiv.org/abs/1708.03985

[9] J. Susskind, G. Hinton, J. Movellan, and A. Anderson, *Generating Facial Expressions with Deep Belief Nets*, 05 2008.

[10] Y. Huang and S. Khan, "Dyadgan: Generating facial expressions in dyadic interactions," 07 2017, pp. 2259–2266.

[11] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," 2017.

[12] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. PP, pp. 1–1, 05 2019.

[13] M. Li, W. Zuo, and D. Zhang, "Deep identity-aware transfer of facial attributes," 10 2016.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 10 2017, pp. 2242–2251.

[15] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," *CoRR*, vol. abs/1611.06355, 2016. [Online]. Available: http://arxiv.org/abs/1611.06355

[16] F. Qiao, N.-M. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Emotional facial expression transfer from a single image via generative adversarial nets," *Computer Animation and Virtual Worlds*, vol. 29, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:46938785

[17] A. Pumarola, A. Agudo, A. M. Martínez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," *CoRR*, vol. abs/1807.09251, 2018. [Online]. Available: http://arxiv.org/abs/1807.09251

[18] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition & Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[19] Y. Yan, Y. Huang, S. Chen, C. Shen, and H. Wang, "Joint deep learning of facial expression synthesis and recognition," *CoRR*, vol. abs/2002.02194, 2020. [Online]. Available: https://arxiv.org/abs/2002.02194

[20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, 04 2016.

[21] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2023, pp. 3138–3147. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCVW60793.2023.00339

[22] H. Jin, S. Liao, and L. Shao, "Pixel-in-pixel net: Towards efficient facial landmark detection in the wild," *International Journal of Computer Vision*, vol. 129, no. 12, p. 3174–3194, Sep. 2021. [Online]. Available: http://dx.doi.org/10.1007/s11263-021-01521-4

[23] B. Ly, E. Dyer, J. Feig, A. Chien, and S. Bino, "Research techniques made simple: Cutaneous colorimetry: A reliable technique for objective skin color measurement," *The Journal of investigative dermatology*, vol. 140, pp. 3–12.e1, 01 2020.

[24] E. Lee, H. Cho, and H. Yoo, "Computational integral imaging reconstruction via elemental image blending without normalization," *Sensors*, vol. 23, p. 5468, 06 2023.