CLC____________ Number__________

UDC____________ Available for reference Yes □No □

Southern University of Science and Technology

Undergraduate Thesis

Thesis Title: **Research on Generated Facial Expression Training Data for Facial Expression Recognition**

Student Name: **Mingjian Zhu**

Student ID: **12012024**

Department: **Department of Computer Science and Engineering**

Program: **Computer Science and Technology**

Thesis Advisor: **Associate Professor Bo Tang**

Date: 6th June, 2024

Commitment of Honesty

1. I solemnly promise that the paper presented comes from my independent research work under my supervisor's supervision. All statistics and images are real and reliable.

2. Except for the annotated reference, the paper contents no other published work or achievement by person or group. All people making important contributions to the thesis of the paper have been indicated clearly in the paper.

3. I promise that I did not plagiarize other people's research achievement or forge related data in the process of designing topic and research content.

4. If there is violation of any intellectual property right, I will take legal responsibility myself.

Signature:

Date: 6th June, 2024

Research on Generated Facial Expression Training Data for Facial Expression Recognition

Mingjian Zhu

(Department of Computer Science and Engineering　Thesis Advisor: Bo Tang)

[**ABSTRACT**]: Facial Expression Recognition (FER) is a key task in computer vision and human-computer interaction with broad applications. However, the performance of FER models is often constrained by the availability and quality of training data. Recently, generative models such as Generative Adversarial Networks (GANs) and Diffusion Models have shown promising results in generating synthetic facial data. This thesis aims to explore the impact of using artificially generated facial expression datasets, along with associated expression labels, on the performance of state-of-the-art FER models. In this work, we first conduct a comprehensive review of the existing literature on FER, focusing on recent advances in addressing challenges such as noisy labels and class imbalance. We then train a state-of-the-art FER model using both real and artificially generated facial expression datasets through full supervision and a state-of-the-art (SOTA) method to evaluate performance. The artificial datasets are constructed using popular generative frameworks, namely GANs and Diffusion Models, and we design a framework to enhance the quality of these generated datasets. Moreover, this work analyzes the performance differences between models trained on real and generated facial expression datasets, discussing the factors contributing to these differences. This is the first work to train FER models using artificially generated data, not only validating the potential of artificial data in addressing the long-tail problem in FER but also providing a new research paradigm for robust FER model training.

[**Keywords**]: Facial Expression Recognition, Long-Tailed Distribution, Synthetic Data, Image Generation

[摘要]：人脸表情识别(FER)是计算机视觉和人机交互中的一项关键任务,应用广泛。然而,FER 模型的性能通常受到训练数据可用性和质量的限制。最近,生成对抗网络和扩散模型等生成模型在生成人脸数据方面显示出了积极的结果。本研究旨在探讨使用这种人工生成的人脸数据集以及相关的表情标签对最先进 FER 模型性能的影响。在这项工作中,首先全面回顾了 FER 方面的现有文献,包括最近在解决噪声标签和类别不平衡等关键挑战方面的进展。本工作基于真实人脸表情数据和人工生成表情数据分别通过全监督学习和一种 SOTA 方法训练一种最先进 FER 模型并进行性能评估。对于人工生成表情数据，分别使用对抗生成网络和扩散模型这两种流行的生成模型框架进行构建，并设计了提高生成数据集质量的调优框架。此外，本工作分析了基于真实和生成人脸表情数据集进行 FER 模型训练的性能差异,并讨论导致这些差异的因素。这是首个基于人工生成数据训练 FER 模型的工作，不仅验证了人工数据在改善 FER 长尾问题的潜力，而且为鲁棒的 FER 模型训练提供了新的研究范式。

[关键词]：面部表情识别；长尾分布；合成数据；图像生成

Table of Content

1. Introduction

1.1 Background and Motivation

Facial expression recognition (FER) is a fundamental task in the field of computer vision and has a wide range of applications, including emotion analysis and detection, human-computer interaction and dialogue systems, security monitoring and behavior analysis, education and training, healthcare, marketing and advertising, and entertainment and gaming. Accurate and robust FER models can enable intelligent systems to better understand human emotions and respond accordingly, leading to more natural, empathetic, and engaging interactions. These models can analyze facial expressions in real-time, allowing systems to detect and interpret the emotional states of users, which can then be used to tailor the system's responses and behaviors to create a more personalized and meaningful experience. By incorporating FER capabilities, intelligent systems can become more attuned to the emotional needs and preferences of their users, fostering stronger connections and improving the overall quality of human-machine interactions.

However, developing effective FER models remains a significant challenge. One of the key obstacles is the long-tailed distribution in existing real-world datasets. Many real-world facial expression datasets exhibit a long-tailed distribution, where certain expressions are significantly underrepresented compared to others. This is primarily because certain expressions are more widely observed and frequently captured in the real world, leading to a significant disparity in the number of samples across different expression categories. Another obstacle is the issue of noisy labels, as the labeling of facial expressions in these datasets can be subjective and prone to errors, especially given the small inter-class distance between certain expressions, which introduces ambiguity in the classification process. Additionally, even though large-scale datasets have been created through significant effort, the samples for some minor expression categories, such as fear, may still be insufficient, making it challenging for models to extract enough knowledge during the training process. Furthermore, the lack of high-quality training datasets is a crucial point that hinders the development of robust FER models. Existing real-world facial expression datasets often lack diversity in terms of factors like age, gender, ethnicity, and cultural background, which can limit the model's ability to generalize to a wide range of users and scenarios. Additionally, the variability in lighting conditions, camera angles, and occlusions present in real-world settings can introduce significant noise and challenges for FER models, making it difficult to achieve accurate and reliable performance. Overall, the development of effective FER models requires addressing these key challenges, including the long-tailed distribution, noisy labels, and the need for more diverse and high-quality training datasets that can capture the nuances of human facial expressions in real-world settings. Overcoming these obstacles is crucial for enabling intelligent systems to truly understand and respond to human emotions in a natural and engaging manner.

These challenges posed by real-world datasets can lead to poor generalization and performance degradation when FER models are deployed in diverse real-world scenarios. The models may struggle to accurately recognize underrepresented expressions, as they have not been exposed to a sufficient number of samples during training to learn the distinctive features of these expressions. Additionally, the models may be overly sensitive to the noise and inconsistencies present in the training data, resulting in suboptimal performance and an inability to reliably interpret the nuanced emotional states of users. Furthermore, the lack of diversity in the training data can also limit the model's ability to generalize to a wide range of users and contexts. If the dataset is skewed towards certain demographic groups or cultural backgrounds, the model may perform poorly when faced with individuals from different backgrounds or with unique facial characteristics. This lack of robustness can undermine the effectiveness of FER-enabled intelligent systems, as they may fail to accurately perceive and respond to the emotional cues of a diverse user base.

To address these challenges and develop FER models that can truly excel in real-world applications, researchers and developers explore innovative approaches to data augmentation, training strategies and robust learning algorithms. This may involve leveraging techniques like mix up, meta learning, long-tailed learning, learning with noisy labels, semi-supervised learning, open-set learning, and cross-domain adaptation to expand the diversity and quality of the training data. Additionally, advancements in model architectures can help mitigate the impact of long-tailed distributions and noisy labels, enabling FER models to achieve more reliable and consistent performance across a wide range of scenarios.

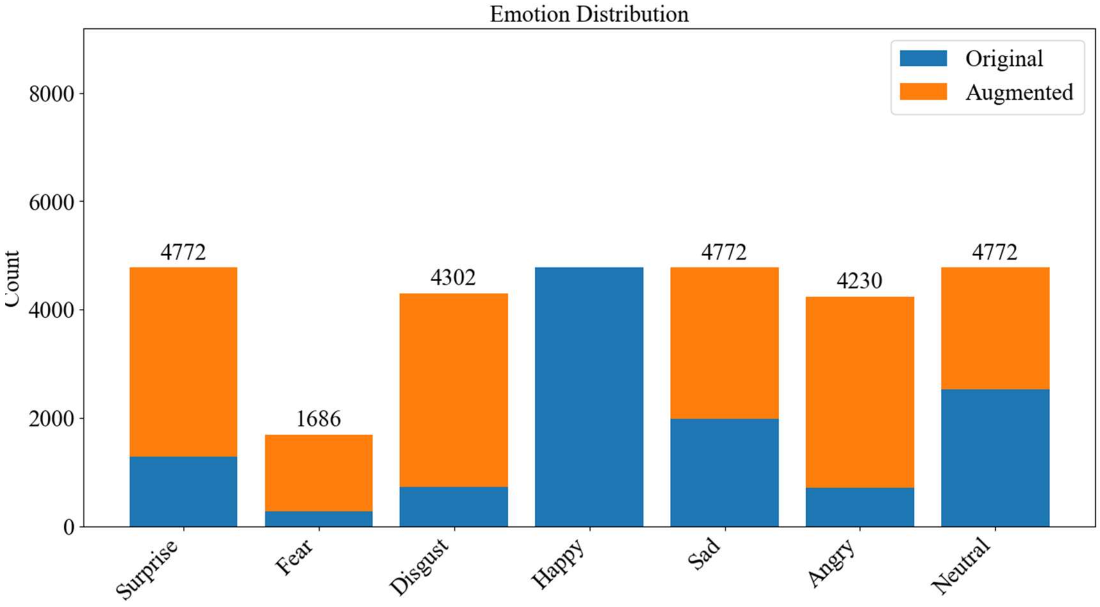


Figure 1 Long-tailed noisy dataset expanded with high-quality generated images for FER

Recently, researchers have explored the use of synthetic data generation techniques, such as generative adversarial networks (GANs) and diffusion models, to augment the training data for recognition tasks like Facial Recognition (FR). Unlike traditional pixel-level transforms, such as flipping, erasing, or mixing up images, these new synthetic data generation methods are more akin to human imagination, allowing for the manipulation of latent space vectors to create novel samples. Newly proposed methods, such as DiscoFaceGAN, DALL-E2, and Stable Diffusion, enable the generation of synthetic data with controlled attributes, enabling the creation of diverse and balanced facial expression samples that can complement the limitations of real-world datasets. These synthetic samples can be annotated with accurate labels, addressing the issue of noisy labeling that often plagues real-world facial expression datasets.

While the use of synthetic data has shown promising results in related domains, such as FR, the application of this approach to Facial Expression Recognition (FER) is still relatively unexplored. Existing studies in FER have not yet extensively investigated the potential of leveraging synthetic facial expression data to improve model performance and robustness. The extent to which synthetic data can effectively substitute or complement real-world data in the context of FER remains an open question that warrants further investigation.

This research aims to address this gap by investigating the impact of incorporating synthetic facial expression data on the performance and robustness of FER models. By conducting a comprehensive evaluation and analysis, the thesis seeks to provide insights into the factors influencing the effectiveness of synthetic data and to identify strategies for optimizing FER systems through the integration of both real-world and synthetic training samples. The findings of this research could contribute to the development of more accurate, robust, and inclusive FER models, enabling intelligent systems to better understand and respond to human emotions in diverse real-world scenarios.

1.2 Problem Statement and Objectives

The primary problem addressed in this research is the limited availability of diverse and representative facial expression datasets, which hinders the development of accurate and robust FER models. Existing real-world datasets often suffer from long-tailed distributions and noisy labels, leading to poor generalization and performance degradation when deployed in diverse real-world scenarios. To address this challenge, researchers have explored the use of synthetic data generation techniques, such as GANs and diffusion models, to augment the training data on relevant domains. These methods allow for the creation of diverse and balanced facial expression samples with controlled attributes, potentially overcoming the limitations of real-world datasets. By generating synthetic data that complements the real-world samples, FER models may be able to learn more robust and generalizable representations of facial expressions. However, while the use of synthetic data has shown promise in related domains, such as Facial Recognition, its effectiveness in the context of FER remains an open question. Factors such as the quality and diversity of the synthetic samples, the optimal integration strategies with real-world data, and the impact on model performance and robustness need to be thoroughly investigated. This research aims to address this gap by conducting a comprehensive evaluation and analysis of the impact of incorporating synthetic facial expression data on the performance and robustness of FER models. The thesis seeks to provide insights into the factors influencing the effectiveness of synthetic data and to identify strategies for optimizing FER systems through the integration of both real-world and synthetic training samples. The findings of this research could contribute to the development of more accurate, robust, and inclusive FER models, enabling intelligent systems to better understand and respond to human emotions in diverse real-world scenarios.

The main objectives of this research are:

Firstly, generate high-quality, accurately labeled, and diverse facial expression images that enable models to acquire sufficient and precise knowledge. Construct balanced facial expression datasets using the aforementioned generated images, strictly ensuring no overlap exists between the generated data and the test data.

Secondly, train facial expression recognition models on these newly created datasets and the original datasets, devising effective evaluation metrics to assess model performance. Compare the performance of models trained on different datasets and analyze the underlying reasons for any discrepancies.

Thirdly, provide insights and guidance for the effective application of synthetic data in the field of facial expression recognition, with the aim of developing more robust and generalizable facial expression recognition systems that can excel in a wide range of real-world scenarios.

By addressing these objectives, this thesis aims to contribute to the advancement of facial expression recognition technology and enable more natural and engaging human-computer interactions.

1.3 Key Contributions

The key contributions of this research are as follows:

Firstly, we implement edge detection as well as comparison among image size to define the quality of images. For low-quality images we implement SwinIR[1] to process them to improve the diversity of generated images. What's more, to improve diversity and accuracy of the labels, we implement GIF[2] to generate images among original datasets, and improve diversity among pose, lighting and shape by finetuning the hyperparameters and prompts. We conduct expensive experiments to generate 5 times large generated data from original RAF-DB[3] dataset with high-quality. We emerge the datasets to compare with original dataset. We strictly ensuring no overlap exists between the generated data and the test data.

Secondly, we demonstrate that our generated images contain enough knowledge for models to learn. We implement the SOTA model Swin Transformer[4], train it on the original dataset and emerged balanced dataset by supervised learning and another SOTA method. We calculate total accuracy, mean accuracy, accuracies of 7 expression classes and confusion matrix on RAF-DB test set and AffectNet[5] test set. We get higher accuracies on minor classes (with over 10% improvement on fear, and improvements on other minor classes) while have little drop among accuracies on major classes (only 1% drop on total accuracy).

Thirdly, we compare images quality generated by GIF and DiscoFaceGAN[6], on the metrices of label correctness, diversity and fidelity. We analyze the gap between generated data and real-world data, provide insights for futural experiments.

Overall, this research makes significant contributions to the field of facial expression recognition and the broader field of computer vision by exploring the potential of synthetic data to address the limitations of real-world datasets and develop more robust and generalizable models.

Here is the article structure. On the following sections, the thesis firstly illustrates related works, including FER techniques and synthetic data generative techniques. The thesis then illustrates method, including model architecture, data preprocessing, datasets, and general framework. The thesis then illustrates experiment, including experiments on GIF generative method and DiscoFaceGAN generative method, trained on RAF-DB training set and augmented RAF-DB training set, using supervised learning and MEK learning method, tested on RAF-DB test set and AffectNet test set. The thesis then gives conclusion.

2. Related Work

2.1 FER Methods

FER is a fundamental task in computer vision and has a wide range of applications, such as human-computer interaction, emotion analysis, and mental health monitoring. Traditional FER approaches often assume that the training and test data follow the same distribution of expression categories[7].

One key challenge in FER is the data bias issue, where existing datasets often exhibit imbalances in factors such as race, gender, and age. This class imbalance problem can lead to poor generalization and suboptimal performance when FER models are deployed in diverse real-world scenarios. To address this challenge, existing FER methods have primarily focused on leveraging labeled facial expression datasets and applying various techniques to mitigate the impact of data biases. However, these approaches are limited by the inherent constraints of the available labeled data.

Orthogonal to these existing methods, the proposed Face2Exp[8] framework aims to enhance FER by utilizing large unlabeled FR datasets. By incorporating the knowledge and diversity present in these FR datasets, the framework has the potential to improve the robustness and generalization of FER models. However, this approach raises another data bias problem – the distribution mismatch between FR and FER data. To combat this mismatch, Face2Exp employs a meta-optimization framework that consists of a base network and an adaptation network. The base network learns prior expression knowledge on class-balanced FER data, while the adaptation network is trained to fit the pseudo-labels of FR data generated by the base model. To further mitigate the impact of the data distribution mismatch, Face2Exp utilizes a circuit feedback mechanism, which improves the base network with the feedback from the adaptation network. This feedback loop helps to align the representations learned by the base network with the characteristics of the FR data, effectively eliminating the data bias between the two domains. Experimental results demonstrate that the proposed Face2Exp framework achieves comparable accuracy to state-of-the-art FER methods, while utilizing only 10% of the labeled FER data used by the baselines. This highlights the effectiveness of the meta-optimization approach in leveraging the complementary information from both labeled FER and unlabeled FR datasets, leading to more robust and inclusive FER models. The Face2Exp framework's ability to combat data biases and improve FER performance across diverse populations represents a significant advancement in the field, paving the way for the development of more accurate and equitable intelligent systems that can better understand and respond to human emotions.

In addition to the data bias challenges, the limited availability of labeled data in minor classes is another obstacle in FER. While there are large-scale datasets that enrich the model training, the class imbalance inherent in these datasets can hinder the learning of robust and generalizable representations. To address this issue, semi-supervised learning approaches have been explored, such as the work by Ada-CM[9]. This method proposes an Adaptive Confidence Margin (Ada-CM) loss function to effectively leverage both labeled and unlabeled data for deep FER.

Unlike traditional semi-supervised learning methods that only select a subset of unlabeled data based on a pre-defined confidence threshold, Ada-CM learns an adaptive confidence margin to make full use of all unlabeled samples. Specifically, Ada-CM partitions the unlabeled data into two subsets: samples with confidence scores higher than the adaptive margin, and samples with confidence scores lower than the margin. For the first subset, the method constrains the predictions to match the pseudo-labels, while the second subset participates in a feature-level contrastive objective to learn more effective facial expression representations. By adaptively leveraging the information from both labeled and unlabeled data, the Ada-CM framework is able to overcome the limitations of class imbalance and data scarcity, leading to state-of-the-art performance in FER. The extensive evaluation on challenging datasets demonstrates the effectiveness of this semi-supervised approach, which can even surpass fully-supervised baselines in certain scenarios. The Ada-CM method represents a significant advancement in addressing the data challenges in FER, complementing the efforts to mitigate data biases. By jointly tackling the issues of class imbalance and limited labeled data, this work contributes to the development of more robust and inclusive

FER models, capable of accurately recognizing facial expressions across diverse real-world scenarios.

Furthermore, real-world FER datasets often contain noisy labels, which can severely degrade the performance of deep neural networks. This challenge is not limited to the FER domain, as the issue of learning from noisy labels is a prevalent problem in numerous deep learning applications. Recent comprehensive surveys[10-11] have examined the various techniques for robust training of deep neural networks in the presence of label noise. These studies categorize the existing methods into two main groups: noise model-based and noise model-free approaches. The former aims to estimate the underlying noise structure and leverage this information to mitigate the adverse effects of noisy labels, while the latter focuses on developing inherently noise-robust algorithms through techniques such as robust loss functions, regularizers, or alternative learning paradigms. By reviewing the state-of-the-art robust training methods across these two categories, the surveys provide a thorough understanding of the current landscape of deep learning with noisy labels. They also delve into the typically used evaluation methodologies, including public noisy datasets and performance metrics, to facilitate the assessment and comparison of these techniques. The insights gained from these comprehensive surveys on learning with noisy labels can serve as a valuable foundation for addressing the challenges posed by imperfect annotations in real-world FER datasets. Adapting and applying the robust training strategies developed for image classification to the FER domain holds the potential to enhance the performance and reliability of deep FER models, enabling their deployment in diverse real-world scenarios.

Another common issue in Facial Expression Recognition (FER) is the class imbalance problem, where certain expression categories are significantly underrepresented in the dataset compared to others. This imbalance poses a significant challenge, as FER models tend to exhibit low performance on the minority expression classes while maintaining high overall accuracy on the test set. To address this challenge, Zhang, Y et al.[12] proposed a novel approach that goes beyond the typical focus on learning knowledge of minor classes solely from minor-class samples.

Inspired by the belief that FER resembles a distribution learning task, where a sample may contain information about multiple classes, their method leverages re-balanced attention maps to regularize the model. This enables the extraction of transformation-invariant information about the minor classes from all training samples, including those from the major classes. Additionally, the authors introduce re-balanced smooth labels to regulate the cross-entropy loss, guiding the model to pay more attention to the minor classes by utilizing the extra information regarding the label distribution of the imbalanced training data. The two proposed modules work together to regularize the model, allowing it to better capture the underlying patterns and features associated with the minority expression classes. Through extensive experiments on different datasets and model backbones, the authors demonstrate that their approach can effectively address the imbalanced FER problem and achieve state-of-the-art performance. Their work highlights the importance of leveraging the inherent information present in the data, even in the major classes, to improve the recognition of underrepresented expression categories in FER.

In real-world FER applications, models may encounter novel expression categories not seen during training, such as compound expressions. To address this, researchers have proposed the Open-Set Facial Expression Recognition[13] (OSFER) task, which allows for unknown expression classes in the test set. While existing open-set recognition methods exist, the authors argue they are not well-suited for OSFER, as FER data have very small inter-class distances, making open-set samples highly similar to closed-set ones. The authors instead propose converting OSFER into a noisy label detection problem, leveraging the sparse distribution of pseudo-labels for open-set samples. The authors introduce a novel method using attention map consistency and cycle training to effectively detect open-set samples. Experiments show this approach outperforms state-of-the-art open-set recognition methods. Their work represents an important advancement in FER, enabling the development of more robust models that can adapt to dynamic human emotional expressions.

The FER research community is benefited from the availability of large-scale real-world datasets, such as AffectNet and RAF-DB, which provide rich annotations for expression categories, as well as demographic information like race and gender. These datasets have enabled the exploration of the aforementioned FER techniques and have driven the advancement of the field.

2.2 Representative Synthetic Generative Methods

Face synthesis has received increasing attention with the great success of GANs. However, generating diversified face images with different expressions, poses, and other attributes has not been well-explored for FER tasks. Some recent works in related fields have explored using synthetic data for model training, providing valuable insights for the FER community. For instance, SynFace[14] proposes a framework to generate high-quality synthetic face images with diverse expressions, poses, and other attributes, and demonstrates its effectiveness in improving FR model performance. Similarly, Expanding Small-Scale Datasets with Guided Imagination (GIF) leverages generative models such as Stable Diffusion[15] to expand small datasets in a fully automatic manner, without requiring manual annotations, and shows promising results in boosting model accuracy on various natural image and medical datasets. While these works are not directly focused on FER, they highlight the potential of synthetic data generation and dataset expansion techniques to address the challenges faced in FER, such as the limited availability of large-scale, high-quality balanced labeled datasets. Exploring similar approaches to generate diverse synthetic face images with rich expression, pose, and attribute annotations could be a promising direction for advancing FER research and overcoming the data scarcity issue. The FER community can draw inspiration from these related works and further investigate the use of synthetic data and automated dataset expansion techniques to enhance the performance and robustness of FER models.

3. Proposed Method

3.1 Overall Framework

The paper presents a comprehensive processing pipeline, as illustrated in Figure 3. The entire workflow is divided into two main parts, image quality check and facial expression check. The following is the proposed overall pipeline of our work.

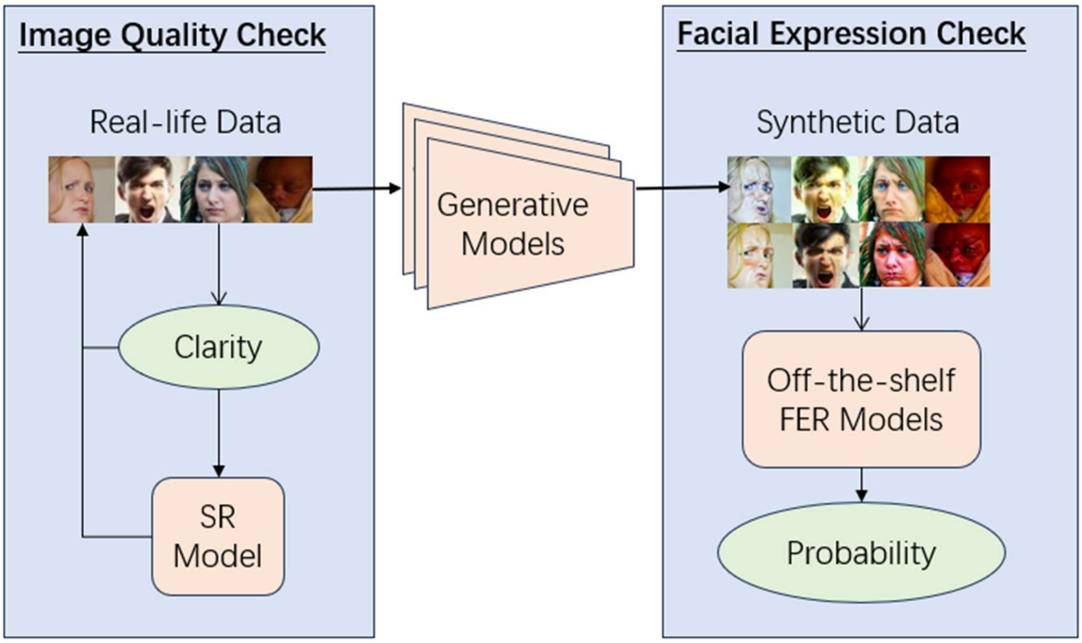


Figure 2 Proposed overall pipeline, pipeline for the best synthetic datasets in our work, the images are from RAF-DB and the synthetic images

The details of several algorithms are illustrated below, noted that we finetune the hypermeters of GIF on RAF-DB dataset, so Generation is finetuned GIF on FER domain, and D is RAF-DB.

The general framework for the thesis is shown in the algorithm, and for more details, some key modules are defined and elaborated below. These modules play crucial roles in the overall system and are essential for understanding the implementation and performance of the proposed method.

Algorithm 1 Pseudo Codes for Framework

Input: A small dataset D, scale s, strength t, time K, prompts, imbalance ratio ρ, generative model G, SR model S, FER recognition model M

Output: Expanded dataset D_{rs}, confusion matrix CM, total accuracy, mean accuracy

```
$D = Quality\_check(D, S)$ $\#check\ out\ the\ low - quality\ images\ and\ enhance\ them$
$D_s = Generation(D, s, t, K, G)$ $\#generate\ synthetic\ images$
$D_{rs} = Augmentation(D, D_S, \rho)$ $\#enhance\ original\ training\ set\ with\ synthetic\ set$
$M(D_{rs})$ $\#\ train\ a\ new\ FER\ model\ with\ added\ synthtic\ data$
```

FER performance has been observed to drop due to the poor quality of existing training data. To address this challenge, it is necessary to conduct a thorough quality check of the dataset. One promising solution to improve FER performance is the use of a powerful super-resolution model called SwinIR (Swin Transformer Iterative Refinement).

SwinIR is a state-of-the-art image super-resolution model that utilizes the Swin Transformer architecture, a versatile and efficient transformer-based model that has shown impressive performance in various computer vision tasks. The key aspects of SwinIR are its ability to effectively capture and leverage the contextual information in images, as well as its efficient design that allows for fast and accurate super-resolution. By applying SwinIR to the low-quality training images, it is possible to enhance the image quality and potentially improve the overall FER performance. In summary, to address the FER performance drop due to poor-quality training data, it is essential to conduct a comprehensive quality check of the dataset. Leveraging the power of the SwinIR super-resolution model can be a valuable approach to enhance the quality of the training data and boost the FER performance.

3.2 Image Quality Check

Algorithm 2, named "Quality_check," is designed to improve the quality of a small dataset D by utilizing a Super-Resolution (SR) model S. The algorithm takes the small dataset D and the SR model S as inputs and outputs an enhanced version of the dataset D with higher quality images.

The algorithm iterates through each image i in the dataset D. For each image, it calculates a quality score using a function called "Calculate_quality_image(i)." This function assesses the quality of the image based on various metrics such as image size, Laplace variance, and other relevant parameters. The specific metrics used can vary depending on the implementation and requirements of the system.

After calculating the quality score for an image, the algorithm compares it against a predefined threshold value. If the quality score is below the threshold, indicating that the image quality is not satisfactory, the algorithm applies the Super-Resolution model S to the image. The SR model S is designed to enhance the resolution and quality of the image, effectively improving its visual characteristics.

The SR model S is SwinIR in our implementation. The model is typically trained on a large dataset of high-quality images and learn to map low-quality or low-resolution images to their high-quality counterparts.

The process continues for each image in the dataset D until all images have been assessed and enhanced if necessary. Finally, the algorithm outputs the updated dataset D, which now contains images of higher quality compared to the original input dataset RAF-DB.

Algorithm 2 Quality_check

```
Input: A small dataset $D$, SR model $S$
Output: A small dataset $D$ with higher quality
for i in images in $D$:
    $score_i = Calculate\_quality\_image(i)$
    #calculate the image size, Laplace variance, etc.
    if $score_i < threshold$:
        $i = S(i)$
```

Algorithm 3, named "Calculate_quality_image," is a function that takes an image i as input and returns a quality score s. The purpose of this algorithm is to assess the quality of an image based on various visual characteristics and metrics. The algorithm combines multiple image analysis techniques to provide a comprehensive evaluation of the image quality.

Algorithm 3 Calculate_quality_image

```
Input: Image $i$
Output: Score $s$

$i_{gray} = ConvertToGrayscale(i)$
$SharpnessScore = ComputeLaplacianSharpness(i_{gray})$
$DetailScore = ComputeFourierDetail(i_{gray})$
$ContrastScore = ComputeContrast(i_{gray})$
$TextureScore = ComputeTextureAnalysis(i_{gray})$
$return\ CombineScores(SharpnessScore, DetailScore, ContrastScore, TextureScore)$
```

3.3 Facial Expression Generation

3.3.1 Synthetic Data Generation

SynFace brings a great approach to generate high quality datasets on facial recognition domain which is relevant with FER, while GIF provides a good way to expand small dataset, creating new data with high diversity and accurate labels by making transforms in latent space points.

Inspired by SynFace, we sample latent vectors from normal distribution. Although leveraging a 3D face reconstruction network[18] to extract the attribute coefficients (expression, illumination, and pose) for each image in the real-world FR dataset is good for futural work, the decoupling areas in encoder and decoder are different in existing works. Using the latent vectors, SynFace then employs a generative model DiscoFaceGAN to synthesize new facial images that have random properties as new synthetic data with different identities, expression, illumination, and pose. The synthetic faces generated in the previous step can be compiled into a new dataset. This dataset has the same statistical properties as real-world datasets because the generative models are also be affected by bias in real-world data. What's more, in FER, we only consider the expression, while the diversity among other factors can make the classifier more robust. Inspired by SynFace, we implement similar way to generate "Syn-FER-80k" datasets, which have similar label distribution and in total 80k images. We then utilizes SOTA models to give labels on the synthetic datasets, implement sets of augmentations and multiple annotations, and by voting we get the labels of the datasets.

GIF guides the imagination of prior generative models based on their criteria. To detail the framework, they use DALL-E2[19] as a prior generative model, which adopts CLIP[20] image/text encoders f_{CLIP-I} and f_{CLIP-} as its image/text encoders and uses a pre-trained diffusion model G as its image decoder. Given a seed image x, they first repeat its latent feature $f = f_{CLIP-I}(x)$ for K times, with K being the expansion ratio. For each latent feature f, they inject perturbation over it with randomly initialized noise $z \sim U(0,1)$ and bias $b \sim N(0,1)$. Here, to prevent out-of-control imagination, they conduct residual multiplicative perturbation on the latent feature f and enforce an ε-ball constraint on the perturbation. Then they follow the proposed criteria to optimize z and b over the latent feature space using meta-learning strategy to boost the class-maintained informativeness and sample diversity of the generated data. Specifically, the class-maintained informativeness score $S_{in\,f}$ encourages the perturbed features to maintain the class semantics of the seed sample while improving their classification informativeness. The sample diversity score $S_{di\,v}$ promotes the diversity among the perturbed latent features. After updating the noise z' and bias b' for each latent feature, they obtain a set of new latent features, which are then used to create new samples through the decoder G. Then a K times larger dataset is created. Inspired by GIF, we expand the training sets of real-world datasets, and then expand the small dataset with K is set to be 5. We then generate new datasets "Syn-FER-Expand", which contains exactly same number of samples for each class with real-world datasets. We also generate balanced datasets using GIF, by sampling balanced data from the real-world training set. For the correctness confirmation, we use the SOTA models to label the synthetic sets and compare to the labels of original data.

By following these approaches, we can create some high-quality synthetic FER datasets that closely match the characteristics of the real-world data, while also allowing for the generation of diverse training samples. These synthetic datasets can then be used to complement the real-world data and evaluate the performance of our FER models in a more controlled and comprehensive manner.

For some low-quality images in real-world datasets such as RAF-DB, which are low in clarity, we also take a quality check to infirm the clarity, and then we utilize Super Resolution (SR) models such as SwinIR to process them. before and after the procession, the label remains correct and the training become more robust.

Algorithm 4, named "Augmentation," is designed to address the issue of class imbalance in a small dataset D by incorporating synthetic data from a synthetic dataset D_s. The algorithm takes the small dataset D, the synthetic dataset D_s, and an imbalance ratio ρ as inputs. The goal is to augment the minor classes in Dataset D with synthetic data to achieve a more balanced class distribution.

Algorithm 4 Augmentation

Input: A small dataset D, synthetic dataset D_s, imbalance ratio ρ

Output: A small dataset D with higher quality

```
$\rho_{og} = num_{max}/num_{min}$
$if\ \rho_{og} > \rho$:
    $join(D, D_s)$ $\#join\ until\ \rho_{og} = \rho\ or\ all\ synthetic\ images\ in\ the\ class\ joined$
    $if\ all\ classes\ checked: break$
```

3.3.2 Synthetic Facial Expression Check

To ensure the quality and correctness of the labels for the synthetic facial expression images, we employ a multi-model approach. This involves using multiple pre-trained models to independently evaluate the synthetic facial expressions and validate the corresponding labels.

The rationale behind this approach is to leverage the diversity and strengths of different models to cross-validate the synthetic data. By utilizing a variety of models, we can establish a more robust and reliable assessment of the synthetic facial expressions, mitigating potential biases or limitations that may be present in a single model.

The models used in this quality check process include multiple emotion recognition models: Deep learning-based models that are capable of classifying the expressed emotions in the synthetic facial images.

By applying these multiple models to the synthetic facial expression data, we can obtain a comprehensive evaluation of the data quality. The models will independently assess the correctness of the facial expressions, the accuracy of the emotion labeling, and the fidelity of the facial features. Any discrepancies or inconsistencies identified through this multi-model assessment will be used to refine and improve the synthetic data, ensuring its reliability for the subsequent FER model training and evaluation.

This rigorous quality check process is a crucial step in ensuring the integrity and usefulness of the synthetic facial expression data, ultimately contributing to the robust performance of the FER models.

3.4 FER Baselines

The FER model serves as a crucial component in the pipeline, encompassing the implementation of various SOTA training frameworks and a comprehensive testing process. Its primary objective is to evaluate the performance of the trained models and provide insights into their effectiveness. Within this module, multiple cutting-edge training frameworks are employed, each designed to optimize the model's learning process and enhance its predictive capabilities.

These frameworks are carefully selected based on their proven track record and ability to handle the specific requirements of the task at hand. Once the models have undergone training using the chosen frameworks, the FER model initiates a rigorous testing phase. During this process, the trained models are applied to test data. The testing process yields several key performance metrics that offer a holistic view of the models' effectiveness. One such metric is the total accuracy, which measures the overall percentage of correct predictions made by the models across the entire test dataset. This provides a high-level indication of the models' predictive power. In addition to total accuracy, the FER model also calculates the mean accuracy. This metric takes into account the performance of the models across different classes or categories, ensuring that the models are not biased towards certain classes while neglecting others. By considering the average accuracy across all classes, the mean accuracy offers a more balanced assessment of the models' performance. Furthermore, the FER model generates a confusion matrix, which is a detailed tabular representation of the models' predictions compared to the actual ground truth labels. The confusion matrix allows for a granular analysis of the models' performance, highlighting the specific instances where the models succeed or struggle in making accurate predictions. It provides valuable insights into the models' strengths and weaknesses, enabling developers to identify areas for improvement and make informed decisions regarding model refinement. By incorporating these comprehensive evaluation metrics, the FER model enables a thorough assessment of the trained models' performance. It not only helps in determining the overall effectiveness of the models but also facilitates the identification of potential issues and guides the iterative process of model enhancement.

In summary, the FER model plays a vital role in the pipeline by implementing SOTA training frameworks and conducting extensive testing to calculate total accuracy, mean accuracy, and confusion matrix. These metrics collectively contribute to a comprehensive understanding of the models' performance and inform the ongoing development and refinement process.

We select models based on their performance on real-world datasets. There are some good common-used backbones in FER task, such as ResNet-18 and ResNet-50[16]. In this work, for better comparison with our selected SOTA method, we employ well-established deep learning architecture Swin Transformer as our baseline models for FER. Swin Transformer is a hierarchical Transformer-based model that uses shifted windows to enable efficient local and global feature learning. This architecture has demonstrated impressive performance on a wide range of visual recognition tasks. The parameters and performances of these baselines are shown in Table 1, we can see that Swin Transformer has relatively higher accuracy.

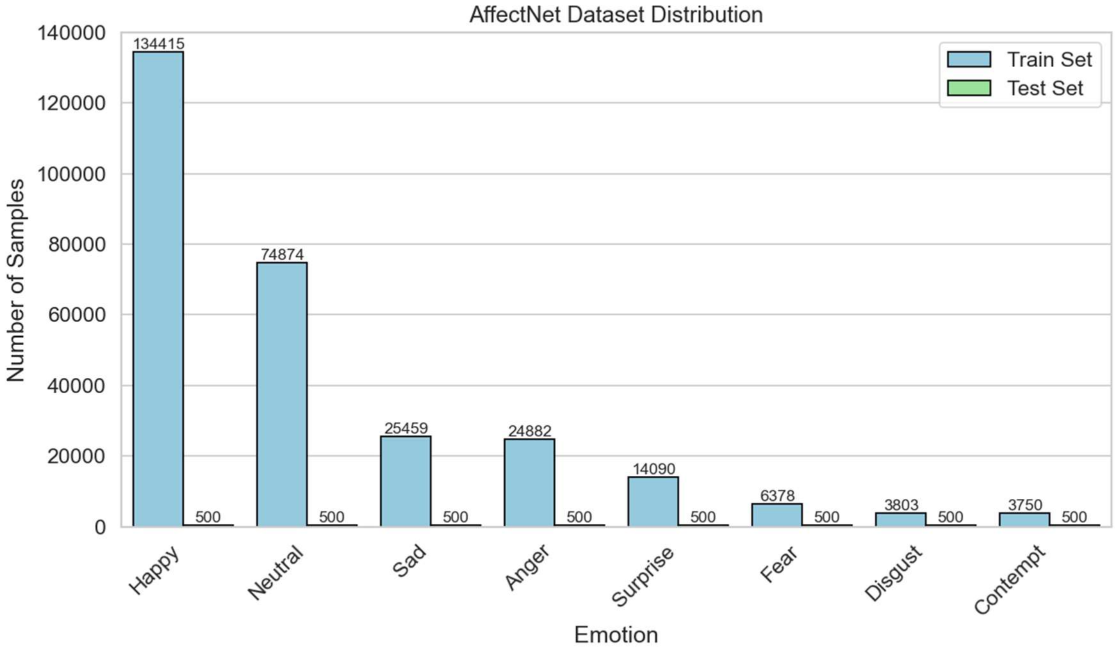
Table 1 Common-used baselines, tested on RAF-DB, accuracies are from experiments conducted by SOTA Leave No Stone Unturned.

Model	# Parameters	Accuracy (%)
ResNet-18	11 M	87.42
ResNet-50	25.6 M	88.33
Swin Transformer	28 M	91.30

We use the pre-trained weights provided by the PyTorch library to initialize the networks. This allows us to leverage the rich visual representations learned from large-scale datasets like ImageNet[17], and then fine-tune the models on the target FER datasets.

The detailed training and evaluation of these baseline models will be discussed in the following sections.

In this work, for real data, we utilize two widely-used benchmark real-world datasets: RAF-DB and AffectNet and generate synthetic datasets from them.



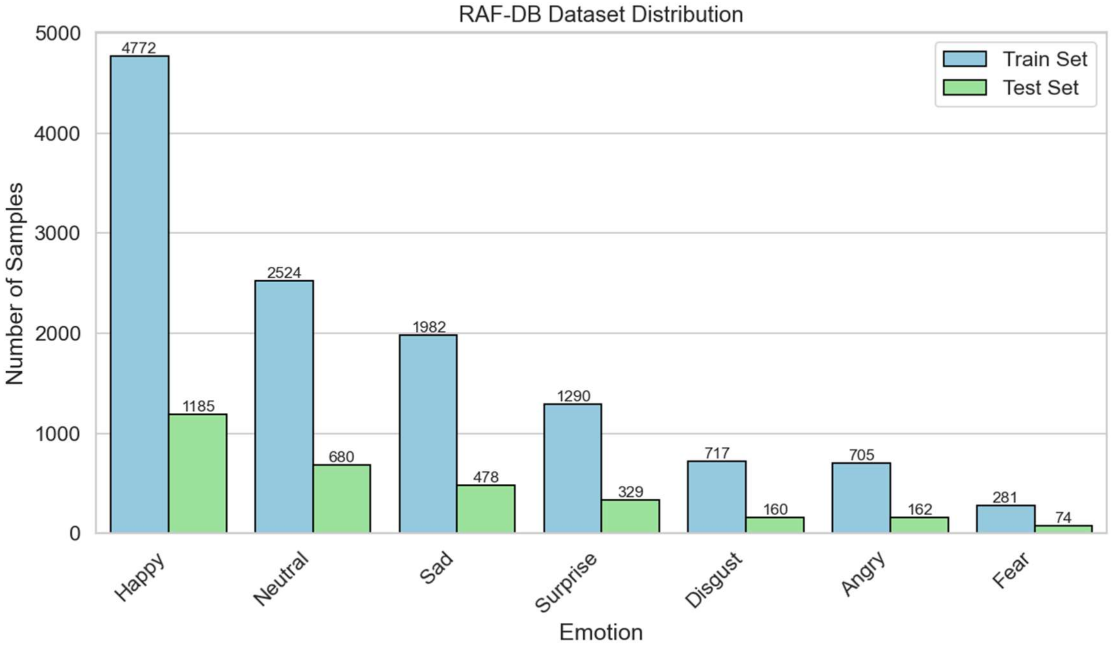


Figure 3 The distribution of different expressions in AffectNet and RAF-DB

The RAF-DB dataset contains 7 basic emotion categories: Surprise, Fear, Disgust, Happy, Sad, Angry, and Neutral. The training set has 1,290 Surprise, 281 Fear, 717 Disgust, 4,772 Happy, 1,982 Sad, 705 Angry, and 2,524 Neutral samples. The test set has 329 Surprise, 74 Fear, 160 Disgust, 1,185 Happy, 478 Sad, 162 Angry, and 680 Neutral samples. The images exhibit large variations in pose, illumination, occlusion, and other real-world factors, making it a challenging dataset for FER.

The AffectNet dataset contains a total of 291,651 facial images, with 75,374 Neutral, 134,915 Happy, 25,959 Sad, 14,590 Surprise, 6,878 Fear, 4,303 Disgust, 25,382 Anger, 4,250 Contempt. To ensure a balanced training process, AffectNet are divided into training set and test set, where 500 images from each class (common practice in relevant works) to create the test set, while the remaining samples were used for training and validation. The training set consists of 291,651 images, and the test set contains 4,000 images, 500 images per classes for 8 classes. The dataset was collected from the internet and covers a diverse range of ages, ethnicities, and real-world imaging conditions.

For preprocessing, we resized all images to a fixed size of 256x256 pixels and performed standard data augmentation techniques, for training samples, we take random crop (randomly crops a portion of the resized image to 224x224), random horizontal flip (flips the image horizontally with a 50% probability), random rotation (randomly rotates the image by up to 5 degrees). For testing samples, we first resize (256x256), and then center crop (crop the center portion of the resized image to 224x224). For all samples, we then normalized the pixel values of all images to the range of [0, 1] by dividing them by 255. The mean and standard deviation of the training set are (0.5863, 0.4595, 0.4030) and (0.2715, 0.2424, 0.2366) (commonly used as standard values for normalizing RGB images in computer vision tasks), respectively.

These datasets and preprocessing ensure that our FER models are trained and evaluated on a diverse and challenging set of facial expressions, which is crucial for assessing their real-world performance. For the synthetic dataset, we implement the same preprocessing.

To complement the real-world dataset, we propose to create a synthetic FER dataset with several key characteristics using existing methods in relevant domains such as SynFace and GIF. First, we aim to generate a large-scale dataset, comparable in class distribution to real-world datasets. This ensures the synthetic dataset has sufficient capacity to train and evaluate FER models effectively and to control the imbalance bias effect.

What's more, we explore two variants of the synthetic dataset - one with a balanced distribution, where each emotion category has an equal number of samples (e.g., 10,000 per class), and another that mimics the imbalanced distribution observed in real-world datasets as mentioned above. This allows us to assess the model's performance under different data skew conditions and develop strategies to handle class imbalance.

Additionally, we will apply same data augmentation techniques to the synthetic dataset, such as flipping, rotating, and scaling the facial images. This helps us understand the impact of data augmentation on model generalization, and how it compares to the effects observed on the real-world data.

Finally, we will carefully monitor the quality and realism of the synthetic data, either through human evaluation or automated image quality assessment metrics (such as predict probability by classifier compared to ground truth). This will ensure the generated samples faithfully capture the characteristics of real facial expressions, leading to more meaningful and insightful experimental results.

By leveraging this versatile synthetic FER dataset, we can conduct comprehensive evaluations of emotion recognition models, paving the way for further improvements and deployments in real-world applications.

4. Experiments

4.1 Experimental Setup

In our experimental setup, we employ a data generation strategy to augment the training set of the RAF-DB dataset. Specifically, we utilize a set of parameters to control the generation process, including K=5, scale=50, and strength=0.1. These parameters are carefully chosen to strike a balance between diversity and quality of the generated samples. By setting K=5, we ensure that each original sample in the RAF-DB training set is used to generate five additional variations. This allows us to significantly expand the training set and expose the model to a wider range of facial expressions and variations. The scale parameter is set to 50, which determines the importance of the prompts applied to the generated samples. This value is selected to introduce meaningful variations while preserving the overall structure and characteristics of the original samples. Furthermore, the strength parameter is set to 0.1, which controls the intensity of the transformations applied to the generated samples. This value is chosen to introduce subtle variations that enhance the model's ability to generalize to unseen data without drastically altering the original samples.

To evaluate the effectiveness of our data generation approach, we test the trained model on both the RAF-DB and AffectNet testing sets. This allows us to assess the model's performance on the original dataset as well as its ability to generalize to a different dataset with potentially different characteristics and challenges. It is worth noting that we also explore other generative models and techniques to further enhance the training data. The details and results of these additional experiments will be discussed in Section 4.3 of our thesis. By investigating various generative approaches, we aim to provide a comprehensive analysis of the impact of data generation on the performance of facial expression recognition models.

This verification module aims to evaluate the performance of a deep learning model for facial expression recognition using the RAF-DB dataset and the expanded dataset. The model is trained using the Adam optimizer with a learning rate of 5e-4 and a weight decay of 0.05. The learning rate schedule follows a cosine annealing strategy, with a warmup learning rate of 5e-7 and a minimum learning rate of 5e-6. The total number of training steps is set to 100,000. The batch size for training is set to 32, and the number of workers for data loading is set to 2. The model is trained with mixed-precision (FP16) to accelerate the training process. The validation process calculates the total accuracy, mean accuracy, and confusion matrix to assess the model's performance. Logging and Checkpointing: The training progress is logged every 1000 steps, and the model states are saved every 6000 steps. The output results and checkpoints are saved in the specified output directory. Reproducibility: The random seed is set to 2048 to ensure reproducibility of the experimental results. The number of classes is set to 7. Label smoothing with a factor of 0.1 is applied.

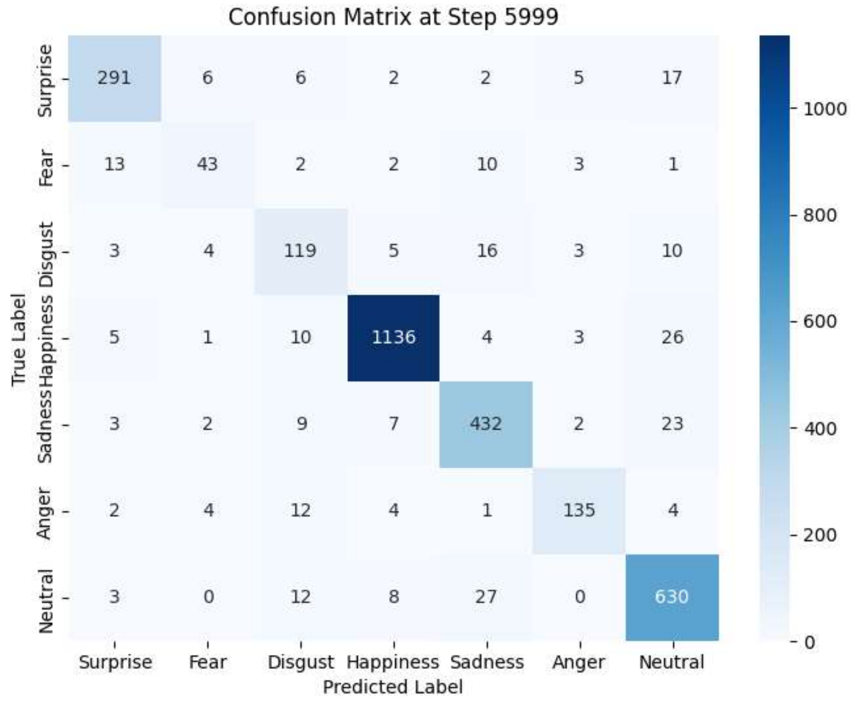
4.2 Results and Discussion

Table 2 Generated by GIF generative method, using Swin-T architecture and MEK learning method, tested on RAF-DB test set, the classes are sorted by the number of samples

Datasets	Happy	Neutral	Sad	Surprise	Disgust	Anger	Fear	Total	Mean
Original (%)	95.86	92.65	90.38	88.45	74.37	83.33	58.11	90.81	83.31
Expanded (%)	**96.54**	91.32	84.73	**88.75**	72.5	**85.19**	**72.97**	90.29	**84.57**

The data generation process employed in our experimental setup yields a diverse set of high-quality images that significantly enhance the training data. By leveraging advanced generative techniques, we are able to introduce a wide range of variations in terms of lighting conditions, hair styles, and other facial attributes. These generated images exhibit a remarkable level of realism and fidelity, making them virtually indistinguishable from the original samples. One of the key strengths of our data generation approach is its ability to maintain the correct label associated with each generated image. Despite the introduction of various transformations and modifications, the generated samples successfully preserve the emotional content and expression label of the original samples. This is crucial for ensuring the integrity and reliability of the training data, as mislabeled or ambiguous samples can potentially mislead the model and hinder its performance.

We can see the confusion matrix, the minor classes in more balanced set share higher accuracies while the total accuracy may has 1 percent lower.



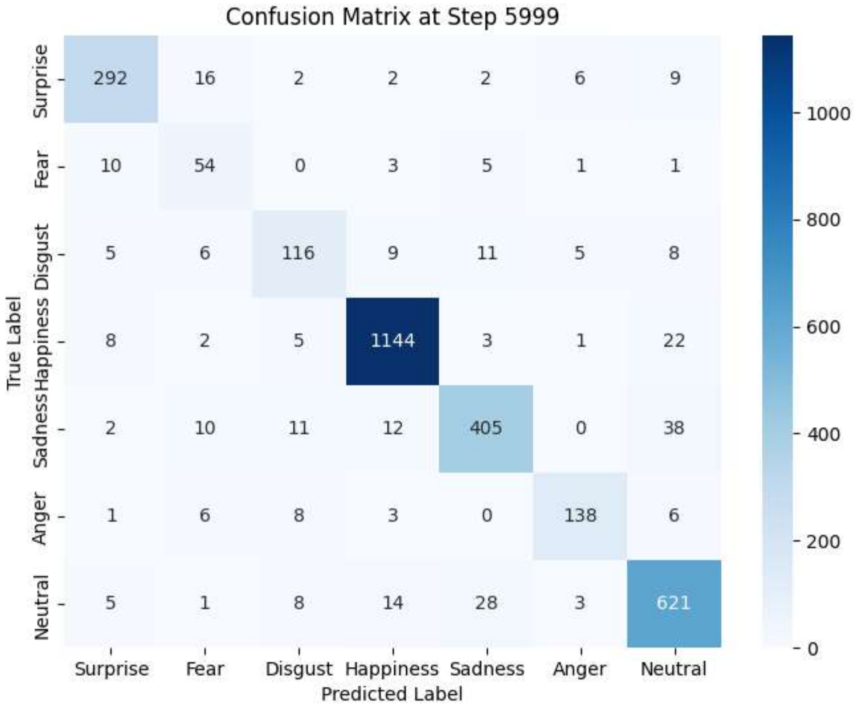


Figure 4　The confusion matrix using the original dataset (the top figure) and the balanced dataset (the bottom figure)

More results will be show below. We conduct expensive experiments to generate high-quality images and undergo lots of failures.

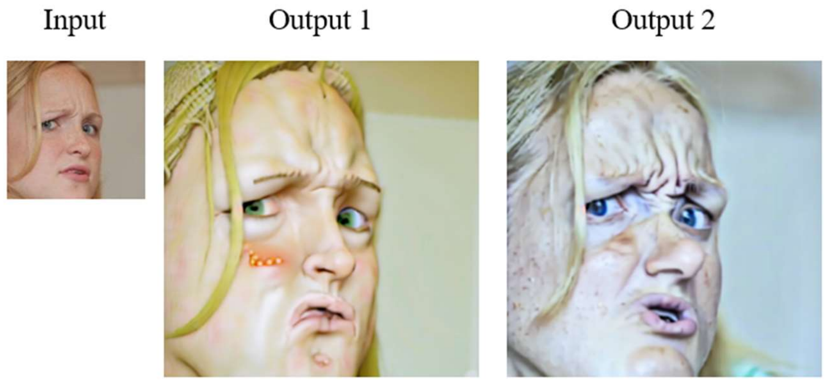


Figure 5　Generation examples for the "disgust" class sample from GIF implemented on RAF-DB

At Table 2, we have total accuracy 90.81% (left) and 90.29% (right), mean accuracy 83.31% (left) and 84.57% (right). Our expanded balanced dataset has much higher mean accuracy with small gap in total accuracy.

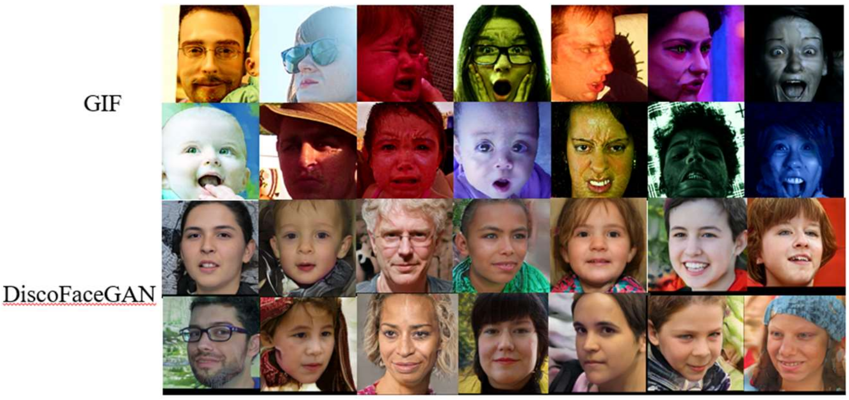


Figure 6　Generation examples of GIF (rows 1, 2) and DiscoFaceGAN (rows 3, 4).From left to right, the emotions in each column are happy, neutral, sad, surprise, disgust, anger and fear

During our analysis of the datasets generated by DiscoFaceGAN, we observed a notable imbalance in the distribution of facial expressions. Upon closer examination, we discovered that a significant majority of the generated images exhibited either a smiling or neutral mouth shape. This skewed distribution raises concerns about the representativeness of the generated dataset and its potential impact on the training of facial expression recognition models.

Table 3 Trained on RAF-DB training set, RAF-DB training set X5 synthetic set, balanced set, using Supervised Learning and MEK tested on RAF-DB test set

Datasets	Happy	Neutral	Sad	Surprise	Disgust	Anger	Fear	Total	Mean
Original_SL (%)	97.38	91.76	89.33	87.54	71.25	88.27	68.92	91.3	84.92
Expanded_SL (%)	96.54	92.79	88.28	90.58	75.63	81.48	70.27	91.26	85.08
Original_MEK (%)	96.96	92.65	89.96	91.79	83.75	87.04	74.32	92.60	88.07
Synthetic_MEK (%)	96.62	94.85	82.22	86.02	59.38	75.93	67.57	89.11	80.37
Expanded_MEK (%)	96.79	92.5	88.91	**92.4**	74.38	**88.89**	72.97	91.98	86.69

The predominance of smiling and neutral expressions in the DiscoFaceGAN-generated dataset suggests that the generative model may have a bias towards these particular mouth shapes. This bias could stem from various factors, such as the training data used to train DiscoFaceGAN or the inherent limitations of the generative model itself. As a result, the generated dataset may not adequately capture the full spectrum of facial expressions, leading to an underrepresentation of other important expressions such as sadness, anger, or surprise.

Table 4 Effect of training Swin-T FER model trained on RAF-DB training set, RAF-DB training set balanced with synthetic set using MEK tested on AffectNet test set

Datasets	Happy	Neutral	Sad	Surprise	Disgust	Anger	Fear	Total	Mean
Original_MEK_Aff (%)	96.2	69.4	44	46.8	24.2	11.6	14.6	43.86	43.86
Expanded_MEK_Aff (%)	93	**77.2**	**44.4**	**48**	**26.4**	7	**15.4**	**44.49**	**44.49**

The imbalance in the distribution of facial expressions in the DiscoFaceGAN-generated dataset can have significant implications for the performance of facial expression recognition models trained on this data. Models trained on a dataset dominated by smiling and neutral expressions may struggle to accurately recognize and classify other expressions when presented with real-world data. This can lead to biased predictions and reduced generalization ability of the trained models.

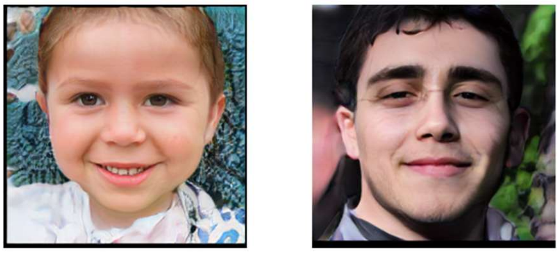


Figure 7 Without filtering, most generated images by DiscoFaceGAN are happy and neutral

Furthermore, our experimental results revealed that the dataset generated using our proposed approach still exhibits a small but noticeable gap in total accuracy compared to the original dataset. Despite the impressive quality and diversity of the generated images, we observed a slight decrease in the accuracies of major classes. This suggests that there may be subtle differences or artifacts introduced during the generation process that impact the model's ability to accurately recognize certain expressions.

Several factors could contribute to this performance gap. One possibility is that the generated images, while visually convincing, may lack some of the fine-grained details or nuances present in the original samples. These subtle differences may be challenging for the model to capture and distinguish, leading to a slight reduction in accuracy. Additionally, the generation process itself may introduce some level of noise or distortion, which could affect the model's ability to extract discriminative features.

To mitigate these issues, further research and refinement of the data generation approach may be necessary. This could involve exploring techniques to ensure a more balanced distribution of facial expressions in the generated dataset, such as incorporating explicit constraints or regularization terms during the generation process.

Additionally, investigating advanced generative models or architectures that can better preserve the subtle details and nuances of facial expressions could help bridge the performance gap.

In conclusion, the imbalance in the distribution of facial expressions in the DiscoFaceGAN-generated dataset and the observed performance gap in our generated dataset highlight the challenges and considerations involved in using generative models for data augmentation in facial expression recognition tasks. Addressing these issues requires careful analysis, iterative refinement, and the development of more sophisticated generative approaches to ensure the generation of diverse, balanced, and high-quality datasets that can effectively support the training of accurate and robust facial expression recognition models.

5. Conclusion

In this thesis, we explored the application of generative models for data augmentation in facial expression recognition tasks. Our experimental setup involved generating additional training samples using a set of carefully chosen parameters to control the generation process. The generated images exhibited high quality, diverse variations in lighting, hair styles, and other facial attributes, while successfully preserving the emotional content and expression labels of the original samples.

However, our analysis of the datasets generated by DiscoFaceGAN revealed a significant imbalance in the distribution of facial expressions, with a predominance of smiling and neutral mouth shapes. This imbalance raises concerns about the representativeness of the generated dataset and its potential impact on the training of facial expression recognition models. Models trained on such biased datasets may struggle to accurately recognize and classify other expressions when presented with real-world data.

Furthermore, our experimental results showed that the dataset generated using our proposed approach still exhibited a small but noticeable gap in total accuracy compared to the original dataset. This suggests that there may be subtle differences or artifacts introduced during the generation process that impact the model's ability to accurately recognize certain expressions.

To address these challenges, future research should focus on developing more advanced generative models and techniques that can ensure a balanced distribution of facial expressions in the generated datasets. This may involve incorporating explicit constraints or regularization terms during the generation process to prevent biases towards specific expressions. Additionally, investigating generative architectures that can better preserve the fine-grained details and nuances of facial expressions could help bridge the performance gap between generated and original datasets.

Reference

[1] Liang J, Cao J, Sun G, et al. Swinir: Image restoration using swin transformer[A]. 2009 IEEE Conference on Computer Vision and Pattern Recognition[C], Ieee, 2021: 1833-1844.

[2] Zhang Y, Zhou D, Hooi B, et al. Expanding small-scale datasets with guided imagination[J]. Advances in Neural Information Processing Systems, 2024, 36.

[3] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[A]. Proceedings of the IEEE conference on computer vision and pattern recognition[C], 2017: 2852-2861.

[4] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[A]. Proceedings of the IEEE/CVF international conference on computer vision[C], 2021: 10012-10022.

[5] Mollahosseini A, Hasani B, Mahoor M H. Affectnet: A database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31.

[6] Deng Y, Yang J, Chen D, Wen F, Tong X. Disentangled and controllable face image generation via 3D imitative-contrastive learning[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C], 2020: 5154-5163.

[7] Canal F Z, Müller T R, Matias J C, et al. A survey on facial emotion recognition techniques: A state-of-the-art literature review[J]. Information Sciences, 2022, 582: 593-617.

[8] Zeng D, Lin Z, Yan X, et al. Face2exp: Combating data biases for facial expression recognition[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C], 2022: 20291-20300.

[9] Li H, Wang N, Yang X, et al. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin[A]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition[C], 2022: 4166-4175.

[10] Algan G, Ulusoy I. Image classification with deep learning in the presence of noisy labels: A survey[J]. Knowledge-Based Systems, 2021, 215: 106771.

[11] Song H, Kim M, Park D, et al. Learning from noisy labels with deep neural networks: A survey[J]. IEEE transactions on neural networks and learning systems, 2022.

[12] Zhang Y, Li Y, Liu X, et al. Leave No Stone Unturned: Mine Extra Knowledge for Imbalanced Facial Expression Recognition[J]. Advances in Neural Information Processing Systems, 2024, 36.

[13] Zhang Y, Yao Y, Liu X, et al. Open-Set Facial Expression Recognition[J]. arXiv preprint arXiv:2401.12507, 2024.

[14] Qiu H, Yu B, Gong D, et al. SynFace: Face recognition with synthetic data[A]. Proceedings of the IEEE/CVF International Conference on Computer Vision[C], 2021: 10880-10890.

[15] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models[A], Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C], 2022: 10684-10695.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[A]. Proceedings of the IEEE conference on computer vision and pattern recognition[C], 2016: 770-778.

[17] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[A]. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE[C], 2009: 248-255.

[18] Deng Y, Yang J, Xu S, et al. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops[C], 2019: 0-0.

[19] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint arXiv:2204.06125, 2022, 1(2): 3.

[20] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[A]. International Conference on Machine Learning[C], PMLR, 2021: 8748-8763.

Acknowledgement

I would like to express my deepest gratitude to my parents, grandmother, sister, teachers, and girlfriend for their unwavering support throughout my undergraduate journey and the completion of this thesis.

To my parents, thank you for your unconditional love, guidance, and the sacrifices you have made to provide me with the best education possible.

To my grandmother and sister, your encouragement and belief in me have been a constant source of motivation.

To my teachers, I am grateful for your invaluable knowledge, patience, and dedication in shaping my academic growth.

Finally, to my girlfriend, your love, understanding, and companionship have been my pillar of strength during challenging times.

This accomplishment would not have been possible without each and every one of you. Thank you from the bottom of my heart.