SYNFER: TOWARDS BOOSTING FACIAL EXPRESSION RECOGNITION WITH SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial expression datasets remain limited in scale due to privacy concerns, the subjectivity of annotations, and the labor-intensive nature of data collection. This limitation poses a significant challenge for developing modern deep learning-based facial expression analysis models, particularly foundation models, that rely on large-scale data for optimal performance. To tackle the overarching and complex challenge, we introduce SynFER (Synthesis of Facial Expressions with Refined Control), a novel framework for synthesizing facial expression image data based on high-level textual descriptions as well as more fine-grained and precise control through facial action units. To ensure the quality and reliability of the synthetic data, we propose a semantic guidance technique to steer the generation process and a pseudo-label generator to help rectify the facial expression labels for the synthetic images. To demonstrate the generation fidelity and the effectiveness of the synthetic data from SynFER, we conduct extensive experiments on representation learning using both synthetic data and real-world data. Experiment results validate the efficacy of the proposed approach and the synthetic data. Notably, our approach achieves a 67.23% classification accuracy on AffectNet when training solely with synthetic data equivalent to the AffectNet training set size, which increases to 69.84% when scaling up to five times the original size. Our code will be made publicly available.

1 INTRODUCTION

Facial Expression Recognition (FER) is at the forefront of advancing AI's ability to interpret human emotions, opening new frontiers for various human-centered applications. From automatic emotion detection to early interventions in mental health Ringeval et al. (2019), accurate pain assessment Huang et al. (2024), and enhancing human-computer interaction Abdat et al. (2011), the potential impact of FER systems is profound Moin et al. (2023); Sajjad et al. (2023); Zhu & Luo (2023). In recent years, learning-based FER models have gained significant traction due to their promising performances Li & Deng (2020); Zhang et al. (2021); Farzaneh & Qi (2021). However, despite recent advancements in network architectures and learning methodologies, the progress of existing FER models has been hindered by the inadequate scale and quality of available training data, underscoring the need to expand datasets with high-quality data to push the boundaries of FER capabilities.

Existing FER datasets, such as CK+ (953 sequences) Lucey et al. (2010), FER-2013 (30,000 48×48 images) Barsoum et al. (2016), RAF-DB (29,672 images) Li et al. (2017), AFEW (113,355 images) Dhall et al. (2017), and SFEW (1,766 images) Dhall et al. (2011), are small compared to popular image datasets for general image processing (e.g., ImageNet Deng et al. (2009) with 1.4 million images and Laion Schuhmann et al. (2022) with billion-level data). While AffectNet Mollahosseini et al. (2017) compiles a large number of facial images from the web, it still suffers from vital drawbacks. A considerable portion of AffectNet's images are low-quality, and its annotations often contain incorrect labels, which impairs the training process of FER models Le et al. (2023); Yan et al. (2022). Consequently, the absence of high-quality and large-scale FER datasets has delayed the development of FER foundation models. However, collecting a large-scale FER dataset with high-quality facial images and meticulous annotations is almost an unrealistic endeavor due to substantial financial and time costs, ethical concerns around facial data collection, and limited resources for large-scale acquisition. Additionally, the subjective interpretation of facial expressions results in

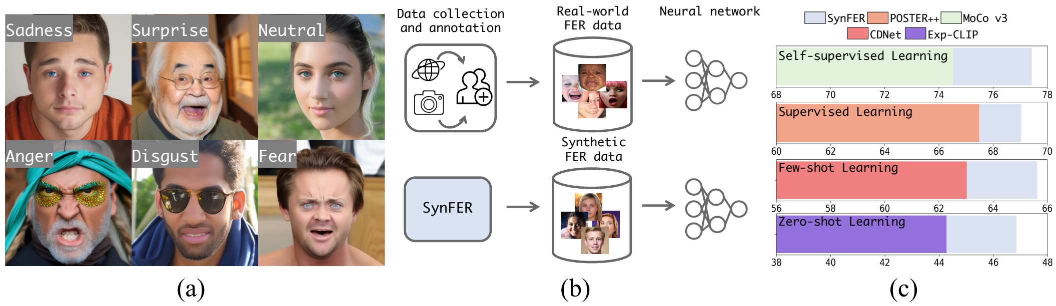


Figure 1: (a) Examples of synthetic facial expression data generated by our SynFER model, (b) Comparison of training paradigms: training with real-world data versus training with synthetic facial expression data and (c) Performance boost from SynFER generated Data in supervised, self-supervised, zero-shot, and few-shot (5-shot) learning tasks.

inconsistent labeling by annotators, which exacerbates variability and hinders the creation of reliable datasets.

To address the challenges in developing FER models, we turn to synthesizing high-quality facial expression images paired with reliable labels. This approach draws inspiration from successful strategies employed to expand annotated datasets for other computer vision tasks, such as semantic segmentation Baranchuk et al. (2021); Chen et al. (2019); Li et al. (2022a) and depth estimation Atapour-Abarghouei & Breckon (2018); Cheng et al. (2020); Guizilini et al. (2022). These advances leverage powerful generative models such as Stable Diffusion Rombach et al. (2022a) and DALL-E Betker et al. (2023), which capture intricate natural image patterns. By tapping into these models, researchers have generated realistic images with their corresponding annotations, thereby boosting model performance. However, applying diffusion models to synthesize facial expression images with reliable FER labels presents two major challenges. (1) the training sets used by these generative models often lack diverse facial expression data, limiting their ability to produce images that capture subtle and nuanced emotional semantics; and (2) prior approaches to generate annotations for synthetic images focused on tangible attributes such as pixel-wise layouts, or depth maps. In contrast, facial expressions convey abstract and subjective emotions, making the generation of precise and reliable expression labels much more complex. To the best of our knowledge, none of the existing methods can simultaneously conduct fine-grained control for facial expression generation and generate robust categorical facial expression labels for face images.

In this paper, we present SynFER, the first framework capable of synthesizing unlimited, diverse and realistic facial expression images paired with reliable expression labels, to drive advancements in FER models. To address the shortcomings of existing FER datasets, which often lack expression-related text paired with facial images, we introduce FEText, a unique hybrid dataset created by curating and filtering data from existing FER and high-quality face datasets. This vision-language dataset serves as the foundation for training our generative model to synthesize facial expression data. To ensure fine-grained control and faithful generation of facial expression images, we inject facial action unit (FAU) information and semantic guidance from external pre-trained FER models. Building upon this, we propose FERAnno, the first diffusion-based label calibrator for FER, which automatically generates reliable annotations for the synthesized images. Together, these innovations position SynFER as a powerful tool for producing large-scale, high-quality facial expression data, offering a significant resource for the development of FER models.

We investigate the effectiveness of the synthetic data across various learning paradigms, demonstrating consistent and modest improvement in model performance. As shown in Fig. 1(c), training with the synthetic data yields significant performance boosts across various learning paradigms. Notably, pre-training on the synthetic data (Fig. 1(b)) with MoCo v3 Chen et al. (2021) yields a significant performance boost of +2.90% on AffectNet, surpassing real-world data pre-training. In supervised learning, SynFER improves accuracy by +1.55% for the state-of-the-art FER model, POSTER++ Mao et al. (2024), on AffectNet. We further explore performance scaling of the synthetic data, revealing further gains as dataset size increases.

Our key contributions are as follows: (1) we introduce FEText, the first dataset of facial expression-related image-text pairs, providing a crucial resource for advancing FER tasks. (2) we propose SynFER, the first diffusion-based data synthesis pipeline for FER, integrating FAUs information and semantic guidance to achieve fine-grained control and faithful expression generation. (3) we develop FERAnno, a novel diffusion-based label calibrator, designed to automatically refine and enhance the annotations of synthesized facial expression images. (4) extensive experiments across various datasets and learning paradigms demonstrate the effectiveness of the proposed SynFER, validating the quality and scalability of its synthesized FER data.

2 RELATED WORK

Facial Expression Recognition (FER): Recent success in deep learning (DL) has largely boosted the performance of the FER task, despite the substantial data requirements for training DL models. To address the limited training data in FER, previous methods mainly focus on developing different learning paradigms, including semi-supervised learning Li et al. (2022b); Yu et al. (2023); Cho et al. (2024), transfer learning Li et al. (2022c); Ruan et al. (2022) and multi-task learning Liu et al. (2023b); Li et al. (2023a). For example, Ada-CM Li et al. (2022b) learns a confidence margin to make full use of the unlabeled facial expression data in a semi-supervised manner. Despite achieving performance gains for FER, these methods remain constrained by limited data. Recently, researchers have explored an alternative data-driven perspective of introducing large-scale face datasets from other facial analysis tasks (e.g., face recognition Zeng et al. (2022)). Meta-Face2Exp Zeng et al. (2022) utilizes large-scale face recognition data to enhance FER by matching the feature distribution between face recognition and FER. However, face data drawn from these datasets lack diverse facial expressions, and thereby couldn't fully unlock the potential of large-scale data in FER.

Synthetic Data: Recently, growing attention has been paid to the advanced generative models (e.g., Generative Adversarial Networks (GANs) Goodfellow et al. (2020) and Diffusion Models Rombach et al. (2022b)), which are typically flexible to synthesize training images for a wider range of downstream tasks, including classification Frid-Adar et al. (2018); Azizi et al. (2023), face recognition Kim et al. (2023); Boutros et al. (2023), semantic segmentation Nguyen et al. (2023); Wu et al. (2024; 2023a) and human pose estimation Feng et al. (2023); Zhou et al. (2023). In particular, some studies pioneer to investigate the capabilities of powerful pre-trained diffusion generative models on natural images Nguyen et al. (2023); Wu et al. (2024); Li et al. (2023b). For example, DatasetDM Wu et al. (2024) further introduces a generalized perception decoder to parse the rich latent space of the pre-trained diffusion model for various downstream tasks. However, there remains a significant gap in research on using diffusion models to generate facial expression data, including both images and corresponding labels. In this paper, we address this gap by exploring diffusion-based synthetic data for the first time in FER. Specifically, we propose to apply a fine-tuned diffusion model to facial expression synthesis and introduce the first diffusion-based pseudo-label generator for FER.

3 PRELIMINARIES

Diffusion models include a forward process that adds Gaussian noise ϵ to convert a clean sample x_0 to noise sample x_T, and a backward process that iteratively performs denoising from x_T to x_0, where T represents the total number of timesteps. The forward process of injecting noise can be formulated as:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon \tag{1}$$

x_t is the noise feature at timestep t and α_t is a predetermined hyperparameter for sampling x_t with a given noise scheduler Song et al. (2020). In the backward process of denoising, given input noise x_t sampled from a Gaussian distribution, a learnable network ϵ_θ estimates the noise at each timestep t with condition c. x_{t-1}, the feature at the previous timestep, is then derived as:

$$x_{t-1} = \frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}} x_t + \sqrt{\alpha_{t-1}} \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \epsilon_\theta(x_t, t, c) \tag{2}$$

During training, the noise estimation network ϵ_θ is guided to conduct denoising with condition c by the learning objective:

$$\min_\theta \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), c, t} \| \epsilon - \epsilon_\theta(x_t, c, t) \|_2^2, \tag{3}$$

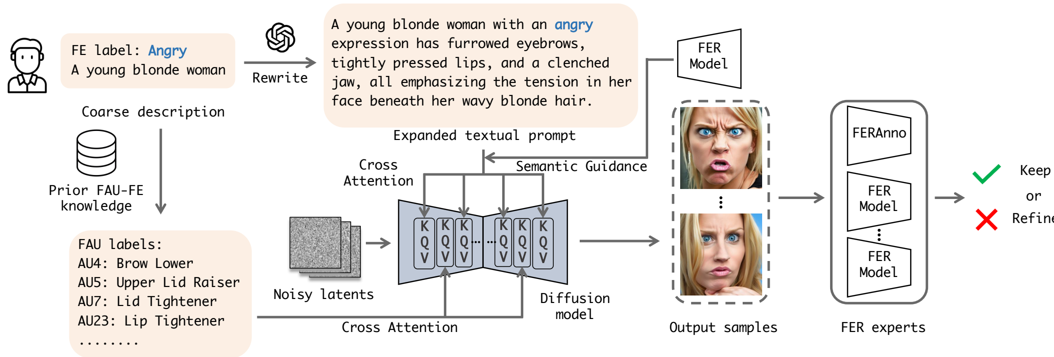


Figure 2: Overall pipeline of our FER data synthesis process.

With its powerful capability to model complex data distributions, the diffusion model serves as the foundation for generating high-quality FER data. Our SynFER framework is the pioneering work that explores the use of diffusion models to synthesize affective modalities.

4 METHODOLOGY

We begin by introducing i) the overall synthetic pipeline for generating facial expression image-label pairs. Next, we detail ii) our approach for producing high-fidelity facial expression images, which are controlled through high-level text descriptions (Sec.4.2.1), fine-grained facial action units corresponding to localized facial muscles (Sec.4.2.3), and a semantic guidance technique (Sec.4.3). Finally, we introduce iii) the FER annotation crafter (FERAnno), a crucial component that thoroughly understands the synthetic facial expression data and automatically generates accurate annotations accordingly(Sec.4.4). This pipeline ensures both precision and reliability in facial expression generation and labeling.

4.1 OVERALL PIPELINE FOR FER DATA SYNTHESIS

We introduce the overall pipeline for FER data synthesis (Fig. 2). The process starts with a coarse human portrait description assigned to a specific facial expression. ChatGPT enriches this description with details such as facial appearance, subtle facial muscle movements, and contextual cues. Simultaneously, facial action unit annotations are generated based on prior FAU-FE knowledge Ekman & Friesen (1978), aligning them with emotion categories to serve as explicit control signals for guiding the facial expression image synthesis. Once the facial expression label, facial action unit labels, and expanded textual prompt are prepared, these inputs condition our diffusion model to generate high-fidelity FER images, guided by semantic guidance to ensure accurate FER semantic. During the denoising process, FERAnno automatically produces pseudo labels for the generated images. To further improve labeling accuracy, we ensemble our FERAnno with existing FER models, which collaborate to vote on the accuracy of the predefined FER labels. In cases where discrepancies arise, the predefined label is refined by averaging the predictions from the ensemble experts. This mechanism effectively reduces the risk of inconsistent or uncertain annotations, ensuring that the final synthesis data is precise and dependable for downstream applications.

4.2 DIFFUSION MODEL TRAINING FOR FER DATA

4.2.1 FETEXT DATA CONSTRUCTION

To address the lack of facial expression image-text pairs for diffusion model training, we introduce FEText (Fig. 3), the first hybrid image-text dataset for FER. It combines face images from FFHQ Karras et al. (2019), CelebA-HQ Karras et al. (2017), AffectNet Mollahosseini et al. (2017) and SFEW Dhall et al. (2011), each paired with captions generated by a multi-modal large language model (MLLM). FEText includes 400,000 curated pairs tailored for facial expression tasks.

Resolution Alignment. Due to variations in image resolution across different datasets, we first utilize a super-resolution model Lin et al. (2023) to standardize the resolutions of images from

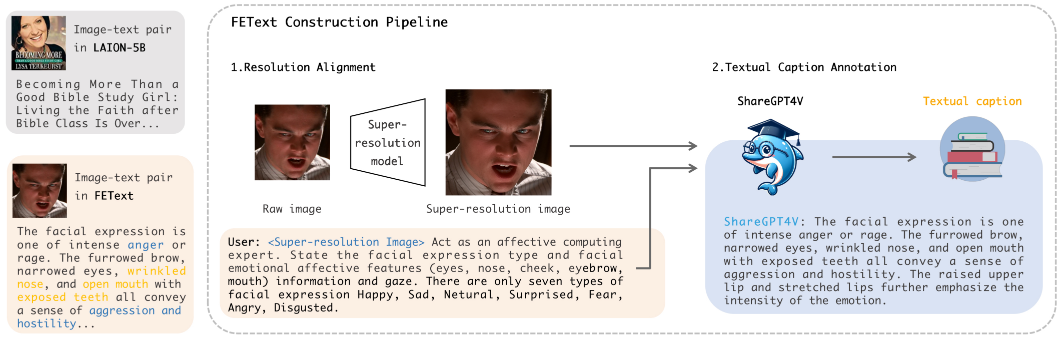


Figure 3: Overview of our FEText data construction pipeline.

AffectNet and SFEW. Specifically, we incorporate high-resolution images from FFHQ and CelebA-HQ datasets to preserve the model's capacity for high-fidelity image generation. This dual approach allows the model to not only maintain the fidelity of the generated images but also to learn and incorporate the facial expression semantics from AffectNet and SFEW.

Textual Caption Annotation. To generate a textural caption for each face image, we employ the open-source multi-modal language model ShareGPT-4V Chen et al. (2023b), by guiding it with carefully crafted instructions. To ensure that the generated captions are both context-aware and expressive, we clearly define the model's role and provide examples of detailed facial expression descriptions within the prompts. This approach enables the model to generate precise, emotion-reflective captions for the input images.

4.2.2 Diffusion Model Fine-tuning

To facilitate our diffusion model to generate high-fidelity facial expressions, a straightforward approach is to fine-tune the model directly on the proposed FEText using the diffusion loss in Eq. 3. However, since FEText contains images processed through a super-resolution model, this direct fine-tuning strategy may lead to over-smoothing in the generated images. To address this, we introduce a two-stage fine-tuning paradigm. In the first stage, the diffusion model is trained on the entire FEText dataset to capture facial expression-related semantics. Then, the second stage mitigates over-smoothing by specifically fine-tuning our diffusion model on the CelebA-HQ and FFHQ subsets of FEText, which consist of native high-resolution images. This two-step approach ensures that our model learns expressive facial details while preserving image sharpness. The fine-tuned model then serves as the foundation for controllable facial expression generation, incorporating facial action unit injection (Sec. 4.2.3) and semantic guidance (Sec. 4.3).

4.2.3 Explicit Control Signals via Facial Action Units

While fine-tuning the diffusion model using facial expression captions provides general language-based guidance for facial expression generation, it lacks the precision needed to capture fine-grained facial details, such as localized muscle movements. To overcome this limitation, we propose to incorporate more explicit control signals through Facial Action Units (FAUs), each of which represents a specific facial muscle movement. Inspired by IP-Adapter Ye et al. (2023), we apply a decoupled cross-attention module to integrate FAU embeddings with the diffusion model's generation process. These embeddings are derived by mapping discrete FAU labels into high-dimensional representations using a Multi-Layer Perceptron, referred to as the AU adapter. FAU labels for each image in the FEText dataset are annotated using the widely adopted FAU detection model, OpenGraphAU Luo et al. (2022). With the diffusion model's parameters frozen, we train the AU adapter to guide the model in recovering facial images based on the annotated FAU labels, using the objective in Eq. 3.

4.3 Semantic Guidance for Precise Expression Control

Due to the imbalanced distribution of FER labels in the training data and the potential ambiguity between certain facial expressions Zhang et al. (2024b), such as *disgust*, relying solely on textual and FAU conditions might not guarantee the faithful generation of these expressions. To address this issue, we propose incorporating semantic guidance on the textual embeddings c^{text}, during the later

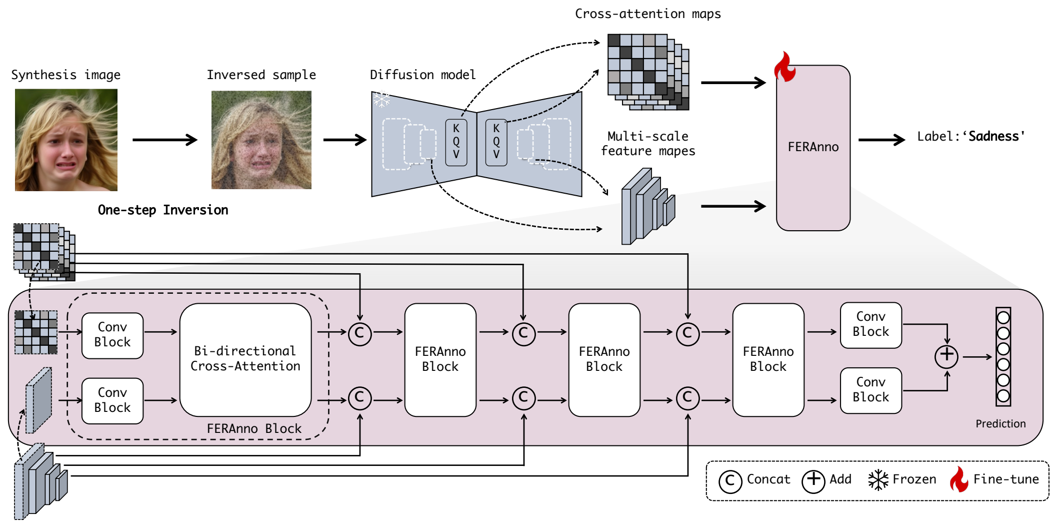


Figure 4: Overview of our FERAnno pseudo-label generator.

stages of the denoising process. We leverage external knowledge from open-source FER models to steer the generation process, ensuring a more accurate and faithful synthesis of hard-to-distinguish facial expressions.

Layout Initialization. During inference, we select a random face image x^s from FEText and invert it to initialize the noise sample x_T^s(Eq. 1). Since early diffusion stages shape the global layout of the image Zhang et al. (2023); Pan et al. (2023); Mao et al. (2023), this strategy helps preserve the natural facial structure, ensuring the generated images are coherent, high-quality, and visually consistent with real-world expressions.

Semantic Guidance. In the early steps of the diffusion process, the generation process is conditioned on the original textual condition c^{text}. To further induce the generation of facial expression images corresponding to their FER labels y, we iteratively update the textual condition in the subsequent time steps. Specifically, a facial expression classifier $f(\cdot)$ is utilized for the injection of complex semantics.

To guide the generated images towards the specific class y, we propose to do so by updating the textual embeddings. Given an intermediate denoised sample x_t at timestep t, following Eq. 15 in DDPM Ho et al. (2020), we first estimate the one-step prediction of the original image $\hat{x}_0$ as:

$$\hat{x}_0 = (x_t - \sqrt{1-\bar{\alpha}_t})\epsilon_\theta(x_t, t, c^{\text{text}}, c^{\text{au}})/\sqrt{\bar{\alpha}_t} \tag{4}$$

We then calculate the classification loss with:

$$\mathcal{L}_g = -y \log(h(f(\hat{x}_0))_i) \tag{5}$$

Given the guidance loss $\mathcal{L}_g$, the textual embedding is updated with the corresponding gradient:

$$c_{t-1}^{\text{text}} = c_t^{\text{text}} + \lambda_g \frac{\nabla_{c_t^{\text{text}}} \mathcal{L}_g}{||\nabla_{c_t^{\text{text}}} \mathcal{L}_g||_2} \tag{6}$$

where λ_g and c_{t-1}^{text} denote the step size and the updated textual embedding at timestep $t-1$, respectively. In the latter steps of the diffusion process, the noise estimator network ϵ_θ is conditioned on the updated textual embeddings rather than the original one.

4.4 Diffusion-based Label Calibrator (FERAnno)

To ensure semantic alignment between each synthesized face image and its assigned facial expression label, we introduce FERAnno, a label calibration framework designed to validate the consistency of the generated data. By analyzing the facial patterns of each synthesized image, FERAnno categorizes them and compares the post-categorized labels with their pre-assigned facial expression labels. This verification process helps identify and filter out samples with mismatched labels, preventing them from negatively impacting downstream FER model training. Specifically, FERAnno is a diffusion-based label calibrator equipped with a deep understanding of facial semantics.

It leverages the multi-scale intermediate features and cross-attention maps inherent in the diffusion model to predict accurate FER labels, as depicted in Fig. 4. This ensures only high-quality, correctly labeled samples are included in the training pipeline, leading to more reliable model performance.

Image Inversion. To extract facial features and cross-attention maps with the diffusion model ϵ_θ, we first inverse the generated image x_0 back to the noise sample x_t at a denoising timestep t, following a predefined scheduler, as described in Eq. 1. To preserve facial details, we set $t = 1$ during the inversion process, ensuring that the facial features remain as close as possible to the original generated image x_0. This partially denoised sample is then passed through the trained denoising network, allowing us to extract rich facial features and cross-attention maps from intermediate layers, which are critical for capturing detailed facial patterns.

Feature Extraction. Given the inverted noise sample x_1 and the corresponding textual condition c^{text} and AU condition c^{au}, we can extract the multi-scale feature representations and textual cross-attention maps from the U-Net ϵ_θ as $\{\mathcal{F}, \mathcal{A}\} = \epsilon_\theta(x_1, t_1, c^{\text{text}}, c^{\text{au}})$, where $\mathcal{F}$ and $\mathcal{A}$ denote the multi-scale feature representations and the cross-attention maps, respectively. $\mathcal{F}$ contains multi-scale feature maps from different layers of the U-Net ϵ_θ with four different resolutions. $\mathcal{A}$ contains the cross-attention maps drawn from the 16 cross-attention blocks in ϵ_θ. Both the feature representation $\mathcal{F}$ and the cross-attention maps $\mathcal{A}$ are regrouped according to their resolutions.

Multi-scale Features and Attention Maps Fusion. Given that the multi-scale feature maps $\mathcal{F}$ capture global information essential for image generation, and the cross-attention maps provide class-discriminative information as well as relationships between object locations Tang et al. (2022); Caron et al. (2021), FERAnno fuses both features and attention maps within a dual-branch encoder architecture for pseudo-label annotation. An overview of this architecture is shown in Fig. 4.

We first compute the mean of the regrouped attention maps, denoted as $\mathcal{A}_{\text{reg}}$, yielding a set of averaged attention maps $\bar{\mathcal{A}}$. Both the feature maps $\mathcal{F}$ and the averaged attention maps $\bar{\mathcal{A}}$ are then passed through a residual convolution block to prepare them for further processing. To effectively integrate information at different scales, we introduce a bi-directional cross-attention block to fuse the features and attention maps. 1×1 convolutions are employed at various stages to adapt the fusion across multiple resolution layers. Finally, the fused feature maps and attention maps are concatenated and passed through a linear layer, which outputs a probability vector for predicting facial expression classes.

5 EXPERIMENTS

We conduct extensive experiments to evaluate both the generation quality of our synthetic data (Sec. 5.1) and its effectiveness in FER tasks (Sec. 5.2). For more details on experimental setup, implementation details are provided in the appendix.

5.1 GENERATION QUALITY

Method	Objective Metrics						User study (Ours vs.)(%)	
	FID (↓)	HPSv2(↑)	FS(↑)	MPS (↑)	FER Acc.(↑)	FAU Acc.(↑)	EA (↑)	FF (↑)
Stable Diffusion	88.40	0.263	2.01	2.00	20.06	87.72	2.86	1.79
PixelArt	145.23	0.271	3.79	5.26	15.52	84.57	24.26	10.00
PlayGround	81.76	0.265	2.86	3.73	21.56	87.28	7.50	5.00
FineFace	74.61	0.268	3.29	1.48	38.05	89.68	5.73	6.41
SynFER	**16.32**	**0.280**	**4.26**	-	**55.14**	**93.31**	**59.64**	**76.79**

Table 1: Ours vs.' shows the proportion of users who prefer our method over the alternative. An MPS above 1.00 and results above 50% in the user study indicate our method outplays the counterpart. FS, FER Acc., FAU Acc., EA and FF denote FaceScore Liao et al. (2024), FER accuracy, facial action unit accuracy, expression alignment and face fidelity, respectively.

We present both objective metrics and subjective user studies, comparing our method to existing state-of-the-art (SOTA) diffusion models Rombach et al. (2022a); Chen et al. (2023a); Li et al. (2024) and the latest facial expression generation technique, FineFace Varanka et al. (2024).

Method	Pre-train Data		RAF-DB	AffectNet	SFEW
	Dataset	Scale			
MCF	Laion-Face	20M	65.22	65.28	32.61
FRA	VGGFace2	3.3M	73.89	57.38	-
PCL	VoxCeleb	1.8M	74.47	68.35	39.68
SimCLR	AffectNet	0.2M	78.65	74.16	46.79
SimCLR	Ours	1.0M	80.24 (+1.59)	75.36 (+1.20)	47.62 (+0.83)
SimCLR	AffectNet+Ours	1.2M	81.52 (+2.87)	75.64 (+1.48)	48.52 (+1.73)
BYOL	AffectNet	0.2M	78.24	73.26	48.70
BYOL	Ours	1.0M	80.96 (+2.72)	75.27 (+2.01)	51.35 (+2.65)
BYOL	AffectNet+Ours	1.2M	81.25 (+3.01)	75.83 (+2.57)	**51.70** (+3.00)
MoCo v3	AffectNet	0.2M	79.05	74.53	49.34
MoCo v3	Ours	1.0M	81.17 (+2.12)	77.43 (+2.90)	50.78 (+1.44)
MoCo v3	AffectNet+Ours	1.2M	**81.68** (+2.63)	**77.82** (+3.29)	51.26 (+1.92)

Table 2: Linear probe performance comparisons of SSL models on three FER datasets.

Method	RAF-DB	AffectNet
ResNet-18	87.48	50.32
ResNet-18 + Ours	87.97	51.65
Ada-DF	90.94	65.34
Ada-DF + Ours	91.21	66.82
POSTER++	91.59	67.49
POSTER++ + Ours	91.95	69.04
APViT	91.78	66.94
APViT + Ours	92.05	67.26
FERAnno	**92.56**	**70.38**

Table 3: Comparison of supervised learning models (with and without our synthetic data) and the label calibrator FERAnno.

Method	CFEE_C		EmotionNet_C		RAF_C	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
InfoPatch	54.19	67.29	48.14	59.84	41.02	57.98
InfoPatch*	55.21	68.73	48.52	61.16	41.88	59.54
LR+DC	53.20	64.18	52.09	60.12	42.90	56.74
LR+DC*	54.65	65.28	51.96	60.14	43.87	57.90
STARTUP	54.89	67.79	52.61	61.95	43.97	59.14
STARTUP*	56.25	69.93	52.87	62.12	45.18	61.23
CDNet	56.99	68.98	55.16	63.03	46.07	63.03
CDNet*	**57.74**	**70.64**	**56.79**	**65.63**	**46.97**	**64.34**

Table 4: Performance comparisons with SOTA few-shot learning methods on 5-way few-shot FER tasks with a 95% confidence interval. (*) indicates training with both real-world data and our synthesis data.

Method	RAF-DB		AffectNet		FERPlus	
	UAR	WAR	UAR	WAR	UAR	WAR
FaRL	24.98	38.53	26.95	26.95	24.27	35.26
FLAVA	14.35	38.69	14.26	14.26	12.50	28.43
CLIP ViT-B/16	38.66	36.16	34.35	34.35	34.33	45.14
Exp-CLIP ViT-B/16	48.96	54.50	39.98	39.98	40.81	53.02
Exp-CLIP* ViT-B/16	51.23	56.32	41.68	41.68	42.33	56.02
CLIP ViT-L/14	47.22	41.13	34.46	34.47	33.82	46.67
Exp-CLIP ViT-L/14	58.70	65.37	44.27	44.27	48.28	55.42
Exp-CLIP* ViT-L/14	**60.41**	**68.34**	**46.85**	**46.85**	**50.26**	**57.98**

Table 5: Zero-shot performance comparison of CLIP for FER, reporting both Weighted Average Recall (WAR) and Unweighted Average Recall (UAR) as in previous works Zhao et al. (2024). (*) indicates models trained with both real-world and synthetic data.

We compute FID between the synthesis images and the test set of the AffectNet Mollahosseini et al. (2017). HPSv2 Wu et al. (2023b) and MPS Zhang et al. (2024a) evaluate the human preferences of the overall synthesis images, while FaceScore Liao et al. (2024) measures the quality of the generated faces. Tab. 1 shows that our method outperforms popular diffusion models and the SOTA facial expression generation method FineFace Varanka et al. (2024), across all metrics of image quality, human preference and facial expression accuracy. Notably, the advantages of SynFER in both FE Acc. and AU Acc. indicate its outstanding controllability in facial expression generation.

5.2 Effectiveness of Synthetic Dataset

Self-supervised Representation Learning. We trained self-supervised learning (SSL) models, including BYOL Grill et al. (2020), MoCo v3 Chen et al. (2021), and SimCLR Chen et al. (2020), using real-world data, our synthetic data, and a combination of both. The linear probe performance of these models was evaluated on three widely used facial expression recognition (FER) datasets: RAF-DB Li et al. (2017), AffectNet Mollahosseini et al. (2017), and SFEW Dhall et al. (2011), with results reported in Tab. 2. All SSL models were trained with a ResNet-50 architecture He et al. (2016). Notably, state-of-the-art methods in self-supervised facial representation learning, such as MCF Wang et al. (2023), FRA Gao & Patras (2024), and PCL Liu et al. (2023c), were pre-trained on much larger face datasets like LAION-Face Zheng et al. (2022), VGGFace2 Cao et al. (2018), and VoxCeleb Nagrani et al. (2020). However, these models underperformed on FER tasks compared to ours, highlighting that existing large-scale face datasets may lack the high-quality and diverse facial expression patterns required for accurate FER. Results demonstrate that combining real-world and synthetic data consistently boosts SSL baselines. Remarkably, even when MoCo v3 was trained solely on our synthetic data, it achieved a 2.12% improvement on RAF-DB, underscoring the effectiveness of our approach in capturing critical facial expression details that are essential for FER.

Supervised Representation Learning. We validate the effectiveness of SynFER for supervised representation learning by evaluating its performance on RAF-DB and AffectNet (Tab. 3). We com-

Figure 5: Generated samples. The first and second rows are fear and disgust, respectively.

Method	HPSv2	FE Acc.	AU Acc.	RAF-DB	AffectNet
Real-world Data	-	-	-	91.59	67.49
SD	0.263	20.06	87.72	89.42	65.36
w/ FEText	0.267	34.62	88.91	90.54	66.62
w/ FEText+FAUs	0.275	48.74	92.37	91.68	67.68
w/ FEText+FAUs+SG	**0.280**	**55.14**	**93.31**	**91.95**	**68.13**

Table 6: Ablation study on the influence of AU injection and semantic guidance (SG) on both the generation quality and supervised representation learning. SD denotes Stable Diffusion , which is used as a baseline.

pare with SOTA FER models, including Ada-DF Liu et al. (2023a), POSTER++ Mao et al. (2024), and APViT Xue et al. (2022). The results demonstrate that incorporating synthetic data consistently enhances both baseline models and the latest SOTAs in supervised facial expression recognition. Notably, APViT benefits from the synthetic data with improvements of 0.27% on RAF-DB and 0.32% on AffectNet. While the improvements in supervised learning are more modest compared to self-supervised learning, they remain consistent. This is likely due to the stricter distribution alignment required in supervised learning between synthetic training data and real-world test data. In the following section on scaling behavior analysis, we provide further insights, showcasing the use of the distribution alignment technique, Real-Fake Yuan et al. (2023), to alleviate this problem.

Few-shot Learning. Addressing the challenge of limited labeled FER data across different scenarios, we explore the potential of synthetic data to enhance few-shot learning, as presented in Tab. 4. Following the protocol established by CDNet Zou et al. (2022), we train models on five basic expression datasets and evaluate them on three compound expression datasets: CFEE_C Du et al. (2014), EmotionNet_C Fabian Benitez-Quiroz et al. (2016), and RAF_C Li et al. (2017). To benchmark our approach, we compare it against SOTA few-shot learning methods, including InfoPatch Liu et al. (2021), LR+DC Yang et al. (2021), and STARTUP Phoo & Hariharan (2021). The results clearly demonstrate that integrating synthetic data consistently enhances few-shot FER performance across key metrics. This highlights the ability of synthetic data, with its broader range of FER patterns, to bridge the gap in data-limited scenarios, allowing models to better generalize to complex, real-world expressions in few-shot tasks.

Multi-modal Fine-tuning. The synthesis data encompasses multiple modalities, including generated images, textual prompts, and FER labels. To assess its impact on multi-modal fine-tuning for FER, we focus on fine-tuning the vision-language foundation model CLIP Wang et al. (2022), as its performance on face-related tasks is widely regarded as sub-optimal Guo et al. (2023); Chen et al. (2024). Building on Exp-CLIP Zhao et al. (2024), we fine-tune the models on CAER-S Lee et al. (2019) and evaluate their zero-shot performances on the datasets outlined in Tab. 5. Our results show that the inclusion of detailed textual prompts and a larger training image set significantly enhances the generalization ability of Exp-CLIP in understanding facial expressions, achieving significant improvements such as +2.58% UAR on the AffectNet dataset.

5.3 Ablation Study

Effectiveness of FAU Control. We validate the effectiveness of SynFER by examining how Facial Action Units (FAUs) enhance the generation process and refine facial details. As illustrated in Fig. 5, samples generated with FAU control (third column) exhibit facial expressions that more accurately match their assigned labels compared to those generated with only text guidance (second column). For example, the ’fear’ expression, driven by FAUs like Inner Brow Raiser and Lip Stretcher, becomes more distinct (third column, second row), making it easier to differentiate from other emotions such as ’surprise.’ Similarly, ’disgust’ is more pronounced with FAUs like Lid Tightener. Without FAU control, facial expressions (second column) tend to blur, as different categories show overlapping features. Quantitative results in Tab. 6 highlight the impact of FAU control: FER accuracy increases from 34.62% to 48.74%, and FAU detection accuracy rises from 88.91% to 92.37%. This also translates into improved downstream performance on RAF-DB and AffectNet.

Effectiveness of Semantic Guidance. We further explore the impact of semantic guidance (SG) on both generation quality and supervised representation learning, as shown in Fig. 5 and Tab. 6. By updating text embeddings to better align with the target facial expression category, SG improves the accuracy of the generated expressions by 6.4%, compared to static text and FAUs. The samples in the last column of Fig. 5 show more exaggerated facial expressions than those in the third column, with SG enhancing the intensity.

Reliability of FERAnno. We assess the reliability of FERAnno as a label calibrator by evaluating its performance on two FER datasets and visualizing its attention maps in Tab. 3 and Fig. 6. FERAnno significantly outperforms previous SOTAs, achieving a +0.51% improvement on RAF-DB and a +1.34% improvement on AffectNet over the second-best models. The attention maps in Fig. 6 further demonstrate FERAnno's ability to accurately locate facial expression-related facial features, such as jaw-dropping and furrowed eyebrows, highlighting the diffusion model's great semantic understanding and fine-grained facial expression recognition.

Synthetic Data Scaling Analysis. Following Tian et al. (2024); Fan et al. (2024), we investigate the scaling behavior of synthetic data in both self-supervised and supervised learning paradigms. To highlight the potential of synthetic FER data, we train models exclusively on synthetic images, without combining real-world data. The results in Fig. 7 (a)-(b) show a stronger scaling effect in self-supervised learning compared to supervised learning, where performance improves significantly with more data. This difference is likely due to the need for better distribution alignment in supervised learning Yuan et al. (2023). While SynFER focuses on addressing FER data scarcity, aligning the synthetic data distribution with real-world data is crucial for supervised tasks. To further explore this, we apply the Real-Fake technique Yuan et al. (2023) for real and synthetic data distribution alignment, and present the results in Fig. 7 (c). Compared to standard supervised learning, Real-Fake demonstrates a clear performance boost.

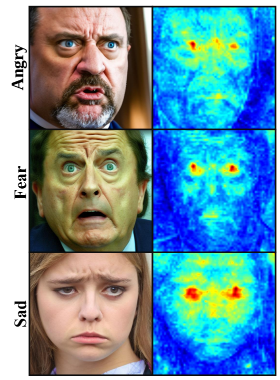
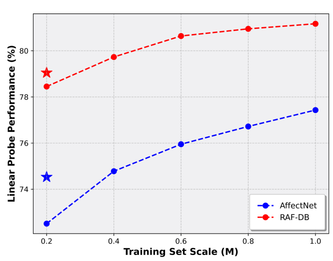
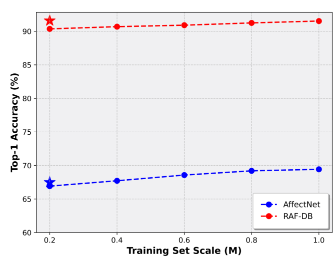


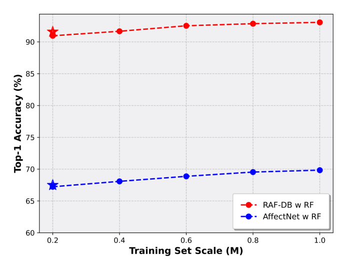
Figure 6: Synthesis images and attention maps in the fine-tuned diffusion model.



(a) Synthetic Data Scaling in Self-Supervised Learning



(b) Synthetic Data Scaling in Supervised Learning



(c) Synthetic Data Scaling in Supervised Learning (with Real-Fake technique)

Figure 7: Scaling up the synthetic FER dataset. MoCo v3 (ResNet-50) Chen et al. (2021) is used for SSL pre-training, and linear probe performance is evaluated on AffectNet and RAF-DB. The SOTA FER model, POSTER++ Mao et al. (2024), is trained using supervised learning (with and without the Real-Fake technique Yuan et al. (2023)) on our synthetic dataset and evaluated on the same two target FER datasets. ★ is model's performance trained on corresponding real data.

6 Conclusion

In this paper, we propose a synthesis data framework SynFER for facial expression recognition to address the data shortage in the field. By consolidating existing FER data and annotating with multi-modal large language models, we introduce the first facial expression-related image-text pair hybrid dataset FEText. We further propose to inject facial action unit information and external knowledge from existing FER models to ensure both fine-grained control and faithful generation of the facial expression images. To incorporate the generated images into training, we propose a diffusion-based label calibrator to help rectify the robust facial expression annotations for the synthesized images. After constructing the data synthesis methodology, we investigate the effectiveness of the synthesis data across different learning paradigms, demonstrating consistent and superior performances. We further study the scaling behavior of the synthesis data for FER.

REFERENCES

Faiza Abdat, Choubeila Maaoui, and Alain Pruski. Human-computer interaction using emotion recognition from facial expression. In *2011 UKSim 5th European Symposium on Computer Modeling and Simulation*, pp. 196–201. IEEE, 2011.

Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2800–2810, 2018.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pp. 279–283, 2016.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19650–19661, 2023.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. *arXiv preprint arXiv:2407.02157*, 2024.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023a.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1841–1850, 2019.

Bin Cheng, Inderjot Singh Saggu, Raunak Shah, Gaurav Bansal, and Dinesh Bharadia. S 3 net: Semantic-aware self-supervised depth estimation with monocular videos and synthetic data. In *European Conference on Computer Vision*, pp. 52–69. Springer, 2020.

Yunseong Cho, Chanwoo Kim, Hoseong Cho, Yunhoe Ku, Eunseo Kim, Muhammadjon Boboev, Joonseok Lee, and Seungryul Baek. Rmfer: Semi-supervised contrastive learning for facial expression recognition with reaction mashup video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5913–5922, 2024.

Victor Guilherme Turrisi Da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. sololearn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pp. 2106–2112. IEEE, 2011.

Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pp. 524–528, 2017.

Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014.

Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5562–5570, 2016.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training ... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7382–7392, June 2024.

Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2402–2411, 2021.

Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14861–14872, 2023.

Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293. IEEE, 2018.

Zheng Gao and Ioannis Patras. Self-supervised facial representation learning with facial region awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2081–2092, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Vitor Guizilini, Kuan-Hui Lee, Rareş Ambruş, and Adrien Gaidon. Learning optical flow, depth, and scene flow without real-world labels. *IEEE Robotics and Automation Letters*, 7(2):3491–3498, 2022.

Jiayi Guo, Hayk Manukyan, Chenyu Yang, Chaofei Wang, Levon Khachatryan, Shant Navasardyan, Shiji Song, Humphrey Shi, and Gao Huang. Faceclip: Facial image-to-video translation via a brief text description. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yuhao Huang, Jay Gopal, Bina Kakusa, Alice H Li, Weichen Huang, Jeffrey B Wang, Amit Persad, Ashwin Ramayya, Josef Parvizi, Vivek P Buch, et al. Naturalistic acute pain states decoded from neural and facial dynamics. *bioRxiv*, 2024.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12715–12725, 2023.

Nhat Le, Khanh Nguyen, Quang Tran, Erman Tjiputra, Bac Le, and Anh Nguyen. Uncertainty-aware label distribution learning for facial expression recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6088–6097, January 2023.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10143–10152, 2019.

Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21330–21340, 2022a.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4166–4175, 2022b.

Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2852–2861, 2017.

Ximan Li, Weihong Deng, Shan Li, and Yong Li. Compound expression recognition in-the-wild with au-assisted meta multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5734–5743, 2023a.

Yingjian Li, Zheng Zhang, Bingzhi Chen, Guangming Lu, and David Zhang. Deep margin-sensitive representation learning for cross-domain facial expression recognition. *IEEE Transactions on Multimedia*, 25:1359–1373, 2022c.

Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. 2023b.

Zhenyi Liao, Qingsong Xie, Chen Chen, Hannan Lu, and Zhijie Deng. Facescore: Benchmarking and enhancing face quality in human generation. 2024. URL https://api.semanticscholar.org/CorpusID:270711518.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8635–8643, 2021.

Shu Liu, Yan Xu, Tongming Wan, and Xiaoyan Kui. A dual-branch adaptive distribution fusion framework for real-world facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Yang Liu, Xingming Zhang, Janne Kauttonen, and Guoying Zhao. Uncertain facial expression recognition via multi-task assisted correction. *IEEE Transactions on Multimedia*, 2023b.

Yuanyuan Liu, Wenbin Wang, Yibing Zhan, Shaoze Feng, Kejun Liu, and Zhe Chen. Pose-disentangled contrastive learning for self-supervised facial representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9717–9728, 2023c.

Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101. IEEE, 2010.

Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 1239–1246, 2022.

Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. Guided image synthesis via initial image editing in diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5321–5329, 2023.

Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, Aibin Huang, and Yigang Wang. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, pp. 110951, 2024.

Anam Moin, Farhan Aadil, Zeeshan Ali, and Dongwann Kang. Emotion recognition framework using multiple modalities for an effective human–computer interaction. *The Journal of Supercomputing*, 79(8):9320–9349, 2023.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.

Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 76872–76892. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f2957e48240c1d90e62b303574871b47-Paper-Conference.pdf.

Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15912–15921, October 2023.

Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=O3Y56aqpChA`.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pp. 3–12, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022b.

Delian Ruan, Rongyun Mo, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. Adaptive deep disturbance-disentangled learning for facial expression recognition. *International Journal of Computer Vision*, 130(2):455–477, 2022.

Muhammad Sajjad, Fath U Min Ullah, Mohib Ullah, Georgia Christodoulou, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad, and Joel JPC Rodrigues. A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. *Alexandria Engineering Journal*, 68:817–840, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024.

Tuomas Varanka, Huai-Qian Khor, Yante Li, Mengting Wei, Hanwei Kung, Nicu Sebe, and Guoying Zhao. Towards localized fine-grained control for facial expression generation. *arXiv preprint arXiv:2407.20175*, 2024.

Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 939–948, 2022.

Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Liang Liu, Yabiao Wang, and Chengjie Wang. Toward high quality facial representation learning. 2023.

Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1206–1217, 2023a.

Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023b.

Fanglei Xue, Qiangchang Wang, Zichang Tan, Zhongsong Ma, and Guodong Guo. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, 14(4):3244–3256, 2022.

Huan Yan, Yu Gu, Xiang Zhang, Yantong Wang, Yusheng Ji, and Fuji Ren. Mitigating label-noise for facial expression recognition in the wild. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022. doi: 10.1109/ICME52920.2022.9859818.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=JWOiYxMG92s`.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Jun Yu, Zhongpeng Cai, Renda Li, Gongpeng Zhao, Guochen Xie, Jichao Zhu, Wangyuan Zhu, Qiang Ling, Lei Wang, Cong Wang, Luyu Qiu, and Wei Zheng. Exploring large-scale unlabeled faces to enhance facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5803–5810, June 2023.

Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. Real-fake: Effective training data synthesis through distribution matching. *arXiv preprint arXiv:2310.10402*, 2023.

Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20291–20300, June 2022.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8018–8027, 2024a.

Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021.

Yuhang Zhang, Yaqi Li, Xuannan Liu, Weihong Deng, et al. Leave no stone unturned: mine extra knowledge for imbalanced facial expression recognition. *Advances in Neural Information Processing Systems*, 36, 2024b.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10146–10156, 2023.

Zengqun Zhao, Yu Cao, Shaogang Gong, and Ioannis Patras. Enhancing zero-shot facial expression recognition by llm knowledge transfer, 2024.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18697–18709, 2022.

Jieming Zhou, Tong Zhang, Zeeshan Hayder, Lars Petersson, and Mehrtash Harandi. Diff3dhpe: A diffusion model for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2092–2102, 2023.

Qihao Zhu and Jianxi Luo. Toward artificial empathy for human-centered design: A framework. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 87318, pp. V03BT03A072. American Society of Mechanical Engineers, 2023.

Xinyi Zou, Yan Yan, Jing-Hao Xue, Si Chen, and Hanzi Wang. Learn-to-decompose: cascaded decomposition network for cross-domain few-shot facial expression recognition. In *European Conference on Computer Vision*, pp. 683–700. Springer, 2022.

A Appendix

A.1 Hyper-parameter Studies

Step Size λ in Semantic Guidance. For hyper-parameter analysis, we consider five configurations of the step size λ in semantic guidance. Due to computational resource constraints, we provide results of self-supervised learning with MoCo v3 Chen et al. (2021) on 0.2M synthetic data for pre-training and report the linear probe performances on RAF-DB Li et al. (2017). Experiment results are shown in Fig. 8. It can be seen that when λ is relatively small, its influence on the performance is relatively small. However, as λ continues to increase, the downstream performance is severely degraded. This is because an excessive λ would lead to severely disrupted images, as shown in Fig. 9.

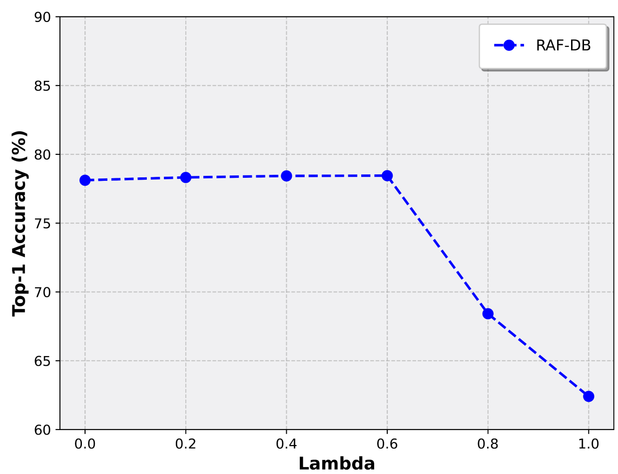


Figure 8: Hyper-parameter study on the λ in semantic guidance. We report the linear probe performances of MoCo v3 Chen et al. (2021) pre-trained with 0.2M synthetic data on RAF-DB Li et al. (2017).

Figure 9: Visualizations of images generated with different values of λ.

A.2 Experiment Setting and Implementation Details

Self-Supervised Learning. We use the widely adopted self-supervised learning library solo-learn Da Costa et al. (2022) for experiments and follow the default settings in solo-learn for various methods. Detailed settings are shown in the tables below:

Others. As all the methods for comparisons in supervised learning (Tab. 3), few-shot learning (Tab. 4) and multi-modal fine-tuning (Tab. 5) are open-source, we thus only need to rewrite the

Config	Pre-Training SimCLR	Pre-Training BYOL	Pre-Training MoCo v3	Linear Probe SimCLR	Linear Probe BYOL	Linear Probe MoCo v3
batch size	64	64	64	32	32	32
optimizer	Lars	Lars	Lars	SGD	SGD	SGD
base learning rate	0.3	0.1	0.3	1e-3	1e-3	1e-3
weight decay	1e-4	1e-6	1e-6	1e-4	1e-4	1e-4
learning rate schedule	warmup cosine	warmup cosine	warmup cosine	step (60,80)	steps (60,80)	steps (60,80)
epochs	200	200	200	100	100	100
augmentation	RRC	RRC	RRC	RRC+RHF	RRC+RHF	RRC+RHF

Table 7: Implementation details on self-supervised pre-training. RRC and RHF denote random resize crop and random horizontal flip, respectively.

corresponding code for dataset reading to incorporate the synthetic data. We follow the default setting in each open-source code of the compared methods.

Facial Action Unit Setting. We use pre-defined facial action unit (FAU) labels to generate images corresponding to specific facial expressions as shown below:

Facial Expression	FAU
Happy	AU6 + AU12
Sad	AU1 + AU4 + AU15
Surprise	AU1 + AU2 + AU5 + AU26
Fear	AU1 + AU2 + AU4 + AU5 + AU7 + AU20 + AU26
Angry	AU4 + AU5 + AU7 + AU23
Disgust	AU9 + AU15 + AU16

Table 8: FAU annotations to generate specific classes of facial expression images.

Step of Performing Semantic Guidance. During the synthesis process, the total denoising steps of the diffusion model are set as 50. Semantic guidance requires backward gradient computation, which would cost a large amount of GPU hours. Thereby, we only perform semantic guidance in the latter steps, which is set as the last 5 steps of the denoising. Another reason to perform semantic guidance at the latter steps is that estimated results at early steps tend to be blurry and degraded facial images, performing semantic guidance on such images might to incorrect results.

Figure 10: Synthetic images comparison between the over-smoothing images and the natural images.

A.3 Over-smoothing of Super Resolution Training Data

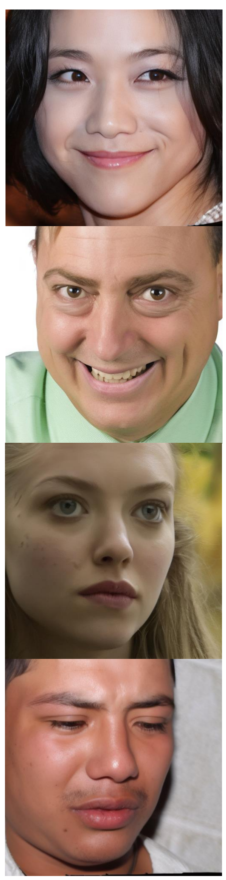
As shown in Fig. 10, we provide comparisons between the over-smoothing synthetic images and the more natural ones. Due to the large amount of super-resolution data in FEText, it can be seen that solely performing fine-tuning on the entire FEText significantly degrades the realism of the images, while the proposed two-stage fine-tuning strategies in Sec. 4.2.2 could effectively prevent over-smoothing.

A.4 Limitations and future work

While the effectiveness of the proposed synthetic data framework has been demonstrated through extensive experiments, its current use is limited to augmenting the training set. A more efficient and optimized approach for leveraging synthetic data remains unexplored and warrants further investigation. Additionally, the generation process remains relatively slow, particularly when incorporating semantic guidance, which is crucial for ensuring accurate and faithful data generation. Moreover, this work focuses exclusively on facial expression recognition. However, it is important to note that the synthetic data framework has potential applications in other areas of facial affective computing, such as facial action unit detection and affective valence and arousal recognition. These avenues are left for future exploration.

A.5 FEText

More examples from FEText are shown in Fig. 11.



The woman in the image is displaying a neutral facial expression. Her eyes are open and focused, and her nose is straight. Her cheeks are slightly puffed out, and her lips are slightly parted, giving her a slight smile. Her eyebrows are relaxed, and her gaze is directed straight ahead. The neutral expression suggests that she is neither happy nor sad, but rather in a state of calm or neutrality.

The man in the image is displaying a Happy facial expression. His eyes are open and looking directly at the camera, indicating that he is engaged and attentive. His nose is straight, and his cheeks are slightly puffed out, adding to the overall cheerfulness of his expression. His mouth is slightly open, and his teeth are visible, which is a common feature of a happy smile. The mans gaze is directed straight at the camera, suggesting that he is comfortable and at ease.

The woman in the image is displaying a neutral facial expression. Her eyes are open, and her gaze is directed to the left. Her nose is straight, and her lips are slightly parted as if she is about to speak. The cheeks of her face are slightly puffed out, and her eyebrows are arched, giving her a thoughtful appearance. The background of the image is blurred, but it appears to be a yellowish-green color, which contrasts with the womans skin tone.

The man in the image is displaying a sad facial expression. His eyes are closed, and his nose is prominent. His cheeks are slightly puffed out, and his eyebrows are furrowed. The mouth is slightly open, and the lips are slightly parted. The gaze is directed downwards, and the overall expression conveys a sense of sadness or melancholy.

Figure 11: Examples from FEText.