SynFER: Towards Boosting Facial Expression Recognition with Synthetic Data

Supplementary Material

1. More Explanation on Method

Fine-tuning Strategy. 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. 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 and semantic guidance.

2. More Experiments

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 [1] on 0.2M synthetic data for pre-training and report the linear probe performances on RAF-DB [4]. Experiment results are shown in Fig. 1. 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. 2.

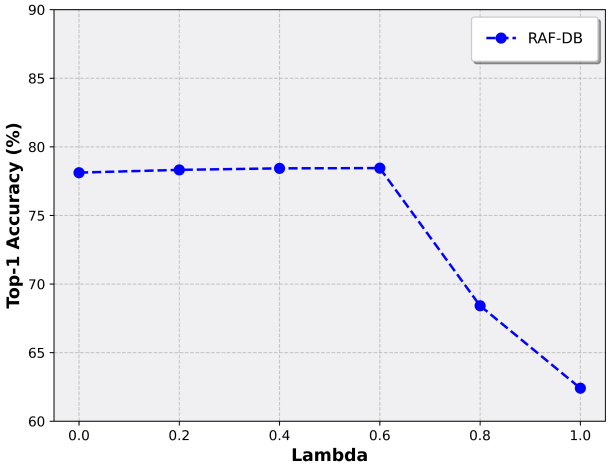

Figure 1. Hyper-parameter study on the λ in semantic guidance. We report the linear probe performances of MoCo v3 [1] pre-trained with 0.2M synthetic data on RAF-DB [4].

Over-smoothing Effect. As shown in Fig. 3, 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 could effectively prevent over-smoothing.

Mitigating Hallucinations of MLLM Annotated Text in FEText. Despite the MLLMs being pretrained on large-scale data, they still have the risk of generating false textual annotations during the curation process of FEText. We validate the effectiveness of our post-processing validation process (Sec. 4.2.1 of main paper) in textual caption annotation. Due to financial constraints, we couldn't manually check all the captions. We randomly sampled 100 image-text pairs, where the captions are generated with or without the validation process, to check their correctness manually. Regarding the expression alignment of the text annotated by MLLM, the validation process could improve the alignment of the annotations from 62% to 81%.

Metadata of the Synthetic Data. In all the evaluations on the effectiveness of the synthetic data generated by SynFER, we randomly sample the FER labels from a unified distribution during the generation process. A similar approach has been conducted to the AU conditions as well. The random sample approach leads to a unified distribution of the FER labels. For the size of the synthetic data, except explicitly stated, we use synthetic data with 5 times the size of the corresponding real-world dataset.

Effectiveness of FERAnno in label voting. In label voting, we use POSTER++ [6], APViT [8] and FERAnno to average the predictions. To validate the effectiveness of FERAnno in label voting, given a set of synthetic images with the size of RAF-DB [4], we refine the labels using two sets of models: (1) POSTER++, APViT and FERAnno, and (2) POSTER++ and APViT. Then, we train ResNet-18 on the label-refined sets. Results are shown in Tab. 1 below, removing the FERAnno from the voting process degrades the trained models' performances by 0.12%, indicating the effectiveness of FERAnno in the voting process. Since it leverages the multi-scale attention maps and feature maps from the diffusion model as inputs, along with the ablation

Voting Models	RAF-DB Acc.(%)
POSTER++, APViT	87.85
POSTER++, APViT, FERAnno	87.97

Table 1. RAF-DB performance comparisons of ResNet-18 models trained on synthetic data

results in Tab. 1, we recognize the FERAnno could have a special view of classifying facial expression compared to traditional networks, thus benefiting the voting process.

3. Experiment Setting and Implementation Details

Self-Supervised Learning. We use the widely adopted self-supervised learning library solo-learn [2] 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, few-shot learning and multi-modal fine-tuning are open-source, we thus only need to rewrite the 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) sets to generate images corresponding to specific facial expressions as shown in Tab. 3. As a specific set of FAU combinations could indicate different classes of facial expressions under various contexts, we randomly sample the AUs in the set for image generation. Notably, during the generation process, both the textual and AU conditions would affect the synthetic output. Therefore, with the sampled AUs ensuring specific local facial movement and text input emphasizing the global facial expression category, the trained diffusion model would automatically complete the rest of the face. This approach could avoid using a static set of AU labels, leading to better facial expression diversity of the synthetic 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.

4. More Discussions

Synthetic Data Diversity. As we highlighted the diversity of the synthetic data in the main paper, instead of demographic inclusivity (i.e. racial diversity), we refer to the generative sampling diversity compared to traditional GAN-based methods, which is widely acknowledged by previous works [3].

Figure 2. Visualizations of images generated with different values of λ.

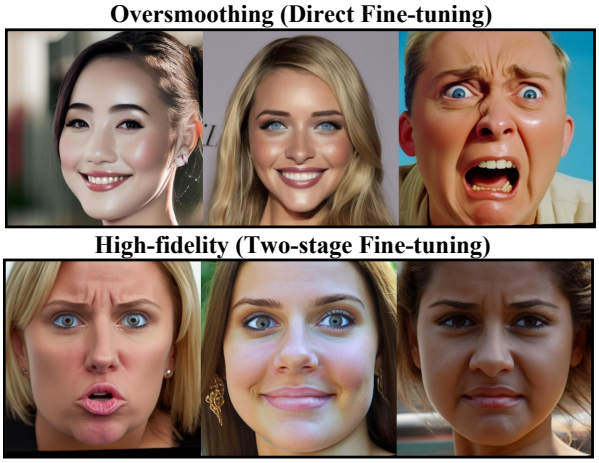

Figure 3. Synthetic images comparison between the oversmoothing images and the natural images.

Inherent Limitations and Empirical Validation of Synthetic FER Data. While this work demonstrates the viability of diffusion-based synthetic data for addressing scarcity in FER training corpora, it is crucial to acknowledge fundamental limitations inherent to all data generation paradigms. Our systematic investigation, spanning human evaluation and empirical experiment results, establishes synthetic data as a functionally viable supplement to constrained real-world datasets. However, it must be noted that neither synthetic nor manually annotated data can guarantee absolute correctness due to irreducible annotation subjectivity and model approximation errors. As empirically validated on multiple datasets and learning paradigms, our approach underscores the practical utility of synthetic data.

Human Evaluation of Expression Alignment and Image Fidelity. In the user study for generation quality evaluation (Tab. 1 of the main paper), we set up a questionnaire asking subjects to choose the images with better expression alignment and generation fidelity between multiple compared methods and ours. Participants were presented with image

Config	Pre-Training			Linear Probe		
	SimCLR	BYOL	MoCo v3	SimCLR	BYOL	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 2. Implementation details on self-supervised pre-training. RRC and RHF denote random resize crop and random horizontal flip, respectively.

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 3. FAU annotations to generate specific classes of facial expression images.

pairs generated by different methods for identical input conditions and asked to select samples demonstrating the best expression alignment and generation fidelity. To calibrate judgments, we provided prototypical examples for each expression category annotated with discriminative features. The study involved 20 participants (predominantly graduate students in computer vision), with responses collected through a double-blind interface to mitigate bias. This protocol aligns with recent benchmarks in multi-modal generative evaluation, where human judgment complements quantitative metrics [5, 7].

5. 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.

6. FEText

More examples from FEText are shown in Fig. 4.

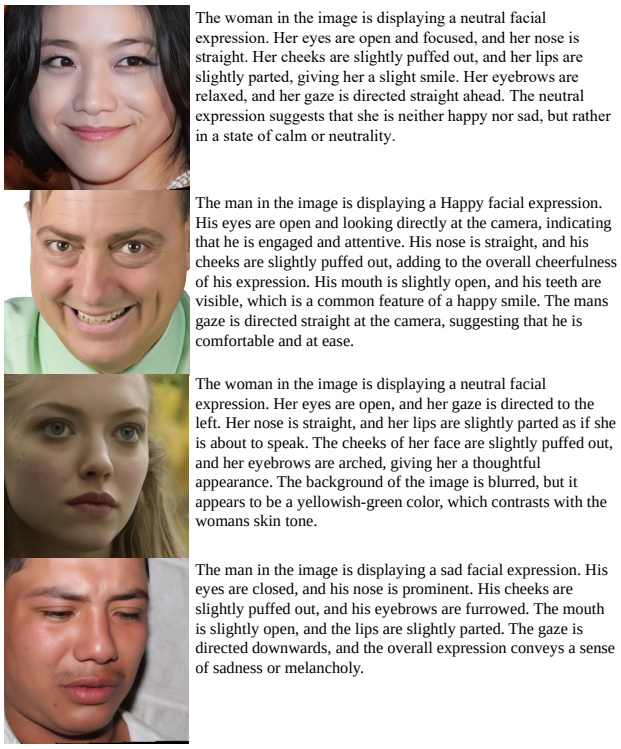

Figure 4. Examples from FEText.

References

[1] 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*, pages 9640–9649, 2021. 1

[2] Victor Guilherme Turrisi Da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022. 2

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[4] 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*, pages 2852–2861, 2017. 1

[5] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[6] 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*, page 110951, 2024. 1

[7] Xiaole Xian, Xilin He, Zenghao Niu, Junliang Zhang, Weicheng Xie, Siyang Song, Zitong Yu, and Linlin Shen. Ca-edit: Causality-aware condition adapter for high-fidelity local facial attribute editing. *arXiv preprint arXiv:2412.13565*, 2024. 3

[8] 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. 1