Response to Reviewer 1 Comments

1. Summary

Thank you for your kind and professional reviews. Here is our point-by-point response to all comments. After thoroughly analyzing the reviewers' feedback, we decided to comprehensively revise the manuscript. Significant corrections were made to the introduction, Section 3.3 on methods, and the experimental section, with minor adjustments to address any non-professional descriptions. All changes have been highlighted in red in the revised submission.

2. Questions for General Evaluation	**Reviewer's Evaluation**	**Response and Revisions**
Does the introduction provide sufficient background and include all relevant references?	Can be improved	Agree and submit after modification.
Are all the cited references relevant to the research?	Can be improved	Agree and submit after modification.
Is the research design appropriate?	Can be improved	Agree and submit after modification.
Are the methods adequately described?	Can be improved	Agree and submit after modification.
Are the results clearly presented?	Yes	Agree
Are the conclusions supported by the results?	Yes	Agree

3. Point-by-point response to Comments and Suggestions for Authors

Comments 1: In Table 2, the latest paper was published in 2019. Enhancing the comprehensiveness of the table by including more recent methods would be beneficial.

Response 1: Thank you for pointing this out. In response, Table 2 presents our usage of four benchmark models: Inception-Resnet-v1[1], Resnet50[2], ConvNeXt[3], and InceptionNeXt[4]. Notably, ConvNeXt[3] and InceptionNeXt[3], introduced in 2022 and 2023 respectively, are state-of-the-art pure convolutional models. Both have demonstrated superior performance in image processing over recent Transformer models. Benefiting from the advancements in ConvNeXt and existing research in the field of face recognition, we conducted studies on a universal model for face recognition. This model demonstrated strong performance across a variety of scenarios. This can be seen in Table 2, lines 367-368 in the article.

Models	Caucasian-test	Indian-test	Asian-test	African-test
Casia [7]	92.15	88.00	83.98	84.93
MS-wo-RFW [7]	93.92	92.98	90.60	90.98
Iman [7]	94.78	94.15	91.15	91.42
Inception-RV1[9]	91.92	86.87	83.35	84.32
ResNet50[10]	92.70	87.40	83.78	84.38
ConvNeXt-T-v1[17]	97.68	93.95	92.10	92.53
InceptionNeXt-T[18]	97.77	93.95	92.28	92.57
FIN-T*	98.23	94.22	93.02	93.22
FIN-S*	98.08	94.75	93.78	93.10
FIN-B*	98.22	95.02	93.72	93.80
FIN-TV2*	98.53	94.85	93.52	93.65
FIN-SV2*	**98.73**	**95.33**	**94.33**	**94.27**

Table 2. Face Fairness Test Set Performance. The first three sets of data in the table come from the paper. The rest of the data are obtained from our experiments.

Comments 2: Transformer-based methods are missing from both Table 2 and Table 3.

Response 2: Thank you for pointing this out. Regarding the issue of not using Transformer models as benchmarks, we offer the following response. Firstly, research using Transformers in the field of face recognition is relatively scarce, and our experiment focuses on combining existing achievements and the latest research in face recognition to propose a new universal model. This model aims to enhance face recognition rates in complex scenarios. Our experiment initially assessed the universal face recognition capabilities of Inception-ResNet-V1[1] and ResNet50[2]. Subsequently, we applied ConvNeXt[3] and InceptionNeXt[4] to the field of face recognition and discovered that these newer models[3,4] exhibited stronger generality than their predecessors[1,2]. By analyzing the performance improvements during our research, we integrated both old and new models to further develop a new model suitable for universal face recognition. Therefore, we chose pure convolutional models as our benchmark and evaluated the effectiveness of our improvements using a complex scenario test set.

As shown in the table below, to provide a clearer understanding of the differences between our method and Transformer performance, we utilized data from the publicly available Face Transformer[5].

Training Data	Models	LFW	SLLFW	CALFW	CPLFW	TALFW	CFP-FP	AgeDB-30
CASIA-WebFace	ResNet-100	99.55	98.65	94.13	90.93	53.17	96.30	95.50
	ViT-P8S8	97.32	90.78	86.78	80.78	83.05	86.60	81.48
	ViT-P12S8	97.42	90.07	87.35	81.60	84.00	85.56	81.48
	FIN-Tiny	**99.22**	**96.72**	**92.13**	**92.48**	**71.42**	**93.39**	**92.07**
	FIN-Tiny-V2	**99.20**	**97.25**	**86.00**	**87.53**	**73.48**	**94.14**	**92.65**
MS-Celeb-1M	ResNet-100	99.82	99.67	96.27	93.43	64.88	96.93	98.27
	ViTP8S8	99.83	99.53	95.92	92.55	74.87	96.19	97.82
	T2TViT	99.82	99.63	95.85	93.00	71.93	96.59	98.07

	ViT-P10S8	99.77	99.63	95.95	92.93	72.95	96.43	97.83
	ViT-P12S8	99.80	99.55	96.18	93.08	70.13	96.77	98.05
MS-wo-RFW	FIN-Tiny	**99.63**	**99.20**	**95.67**	**89.17**	**64.72**	**94.54**	**97.30**
	FIN-Tiny-V2	**99.73**	**99.38**	**95.52**	**90.03**	**66.28**	**95.54**	**97.72**

Comments 3: Although the authors claim that the proposed method can address complex cases, these cases are not detailed in the experimental section.

Response 3: In response to the question, we offer the following: Our primary objective is to design a versatile facial recognition model. Its effectiveness is demonstrated through comparisons with baseline models in complex environments. We revisit our motivation for evaluating the model in complex scenarios in the introduction, lines 59-62. The distinctions among various test sets are detailed in the experimental section, lines 275-284. Additionally, the performance of our approach on specialized scenario test sets is presented in the experimental section, lines 370-380, thereby assessing the model's versatility.

Introduction 59-62: To thoroughly evaluate the effectiveness of our proposed improvements, we introduced verification datasets from multiple scenarios and measured the models' universality by calculating their average precision across these datasets.

Experiments 275-284: Given privacy concerns in face recognition, benchmarks typically encompass 5,000 to 8,000 face matching tasks. We utilized 14 benchmarks for accuracy assessment. These include LFW, which tests face recognition algorithms under real-world conditions, AgeDB-30 for age invariance, CFP-FF and CFP-FP for pose variations, and VGG2-FP for frontal to profile comparisons. For extreme scenarios, we used CALFW to assess real-world variations, CPLFW for cross-pose recognition, MLFW for ethnic diversity, SLLFW for low-light conditions, and TALFW for temporal aging variations. Additionally, RFW provided fairness metrics by evaluating racial fairness across diverse racial groups (Africa-test, Asian-test, Caucasian-test, and Indian-test). AgeDB-30, VGG2-FP, and CALFW were also pivotal in evaluating cross-age recognition performance.

Experiments 370-380: On the face-balanced test set and Special Scenes Test Set. Our findings exceed those reported by IMAN, based on comparative data sourced from their paper. The analysis shows our model, FIN-SV2, significantly surpasses previous models in accuracy while also reducing racial bias. This improvement is particularly notable in test sets for African, Asian, and Indian demographics, where our model achieved accuracies over 0.94. These results, detailed in Table 2, suggest that using large datasets can mitigate racial bias in model performance effectively. Additionally, we assessed the model's accuracy in various Special Scenes, with results shown in ROC curves Figure 5 and Figure 6. The experiments confirm that our improvements enhance facial recognition performance in complex scenarios. An analysis of the difference between optimal and average values further demonstrates the versatility and robustness of our approach.

Comments 4: This paper focuses on the face recognition task, with applications extending to related fields such as expression recognition [1], expression generation [2][3], and face reconstruction [4]. Exploring the relationships between these applications could provide insightful analysis in the introduction section.

[1] Facial Prior Guided Micro-Expression Generation

[2] Template inversion attack against face recognition systems using 3d face reconstruction
[3] Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction

Response 4: Regarding this issue, we offer the following response: While our research focuses primarily on recognition, expanding the underlying model's applications is highly valuable. Therefore, we modify the introduction and add analysis on masked faces [6], age scenes [7] and face reconstruction [8, 9, 10]. For details, please see lines 41-45 of the article.

Introduction 41-45: Zhang et al. [6] improved the Inception module, significantly enhancing performance on datasets featuring masked faces. LIAAD[7] developed a novel, lightweight attention-based method that, through knowledge distillation, improves accuracy and robustness against age variations in face recognition. We delves into face recognition and its applications, including expression recognition[8], generation, and face reconstruction[9,10].

[1]Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).
[2]He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[3]Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.
[4]Yu W, Zhou P, Yan S, et al. Inceptionnext: When inception meets convnext[J]. arXiv preprint arXiv:2303.16900, 2023.
[5]Zhong Y, Deng W. Face transformer for recognition[J]. ar**v preprint ar**v:2103.14803, 2021.
[6] Zhang, Mengya, Yuan Zhang, and Qinghui Zhang. "Attention-Mechanism-Based Models for Unconstrained Face Recognition with Mask Occlusion." Electronics 12.18 (2023): 3916.
[7] Truong, Thanh-Dat, et al. "LIAAD: Lightweight attentive angular distillation for large-scale age-invariant face recognition." Neurocomputing 543 (2023): 126198.
[8] Zhang, Y., Xu, X., Zhao, Y., Wen, Y., Tang, Z., & Liu, M. (2023). Facial prior guided micro-expression generation. IEEE Transactions on Image Processing.
[9] Shahreza, Hatef Otroshi, and Sébastien Marcel. "Template inversion attack against face recognition systems using 3d face reconstruction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[10]Yang, Haotian, et al. "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction." Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2020.

4. Response to Comments on the Quality of English Language

Response 1: We revised the inappropriate sections of the manuscript, highlighting significant revisions in red and stylistic refinements in another shade of red. We thank the reviewers for their constructive feedback.

5. Additional clarifications

None.