2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

Test-Time Augmentation for Pose-invariant Face Recognition

Jaemin Jung*, Youngjoon Jang*, Joon Son Chung
Korea Advanced Institute of Science and Technology, South Korea

***Abstract*— The goal of this paper is to enhance face recognition performance by augmenting head poses during the testing phase. Existing methods often rely on training on frontalised images or learning pose-invariant representations, yet both approaches typically require re-training and testing for each dataset, involving a substantial amount of effort. In contrast, this study proposes Pose-TTA, a novel approach that aligns faces at inference time without additional training.**

To achieve this, we employ a portrait animator that transfers the source image identity into the pose of a driving image. Instead of frontalising a side-profile face – which can introduce distortion – Pose-TTA generates matching side-profile images for comparison, thereby reducing identity information loss. Furthermore, we propose a weighted feature aggregation strategy to address any distortions or biases arising from the synthetic data, thus enhancing the reliability of the augmented images.

Extensive experiments on diverse datasets and with various pre-trained face recognition models demonstrate that Pose-TTA consistently improves inference performance. Moreover, our method is straightforward to integrate into existing face recognition pipelines, as it requires no retraining or fine-tuning of the underlying recognition models.

I. INTRODUCTION

Face recognition has advanced considerably with the development of deep learning technologies [7, 38, 50, 56, 57]. However, early face recognition models were trained without accounting for pose variations, leading to reduced reliability when encountering faces with unseen poses. In response, several studies introduced datasets with diverse head poses [2, 6, 11, 12, 30, 61, 66], laying the foundation for research on pose-agnostic face recognition. Thanks to these studies, the scale of datasets has grown, and models [8, 17, 33, 42, 49] have become more advanced, resulting in face recognition performance exceeding 90% accuracy across various datasets. However, recent works demonstrate that there is still room for improvement in scenarios involving diverse pose variations. In particular, side-profile images present substantial intra-personal variations, making recognition more challenging. Two primary approaches have been explored: (1) training models to extract pose-invariant representations [1, 34, 36] and (2) employing frontalisation techniques to synthesise images for model training [4, 15, 21, 40, 53].

One approach to learning pose-invariant representations is proposed in PoseFace [36], where an orthogonality constraint is applied to separate identity information from head pose in the input image. This constraint allows the model to learn representations in which images of the same identity are mapped to the same embedding space, regardless of the facial pose. Similarly, SSN [34] utilises a symmetrical siamese network with shared weights and contrastive loss to effectively align embeddings of the same identity across various poses. However, as these approaches use a pre-trained head pose extractor, their performance is inherently dependent on the accuracy of this model.

In contrast to these methods, face normalisation approaches aim to directly manipulate the facial pose to align faces into a canonical view. CAPG-GAN [21] performs generative model-based face frontalisation by using the target pose as input to generate faces with the desired head pose. However, converting a side-profile image to a frontal view often results in information loss, affecting features such as facial hair, wrinkles, and overall face shape. To mitigate this issue, Dual-Attention GAN [62] focuses on key facial features like eyes, nose, and mouth while considering facial symmetry, and High-Fidelity Face Manipulation [16] uses a high-resolution GAN with an attention mechanism to preserve facial details. However, a major limitation of existing works is that they all require additional training processes, and validating their generalisation requires substantial computational resources and extensive experiments.

In this paper, we propose a method to enhance the performance of a pre-trained face recognition model during the inference process using test-time augmentation (TTA), without the need for any additional training steps. Unlike existing frameworks, we reduce the dependency on pose estimators by adapting a portrait animator [9, 13, 18, 23, 46, 47, 58, 60] that takes a driving image and a source image as input, generating an image that imitates the driving pose while maintaining the identity of the source image. Next, we focus on the issue of face distortion that occurs when converting side-profile images to frontal ones. Rather than performing face verification on a distorted frontal face, we propose a more effective approach with our method, Pose-TTA. During the pose augmentation process, instead of generating a frontal face from the unseen side of the face, Pose-TTA modifies the given face to a side profile and aligns the two images for comparison. This method minimises the loss of identity information by comparing two spatially aligned faces, thereby providing indirect supervision for the regions of the face on which the model should focus.

Additionally, unlike traditional TTA methods [10, 25, 29, 39, 44, 45] that do not use generative models and maintain object information within an image, Pose-TTA leverages a generative model to synthesise the target pose. To address potential biases and distortions in the synthetic data, we propose a weighted feature aggregation approach based on the reliability of the synthetic data. This approach ensures that the TTA process accounts for any inconsistencies in

*These authors contributed equally to this work.

979-8-3315-5341-8/25/$31.00 ©2025 IEEE

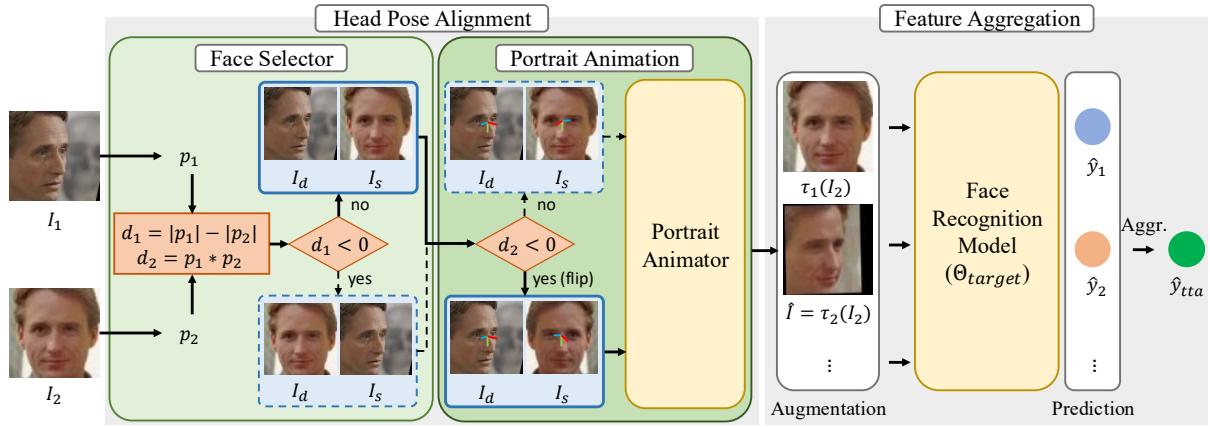

Fig. 1. An overview of our Pose-TTA framework. In the head pose alignment stage, the source and driving images, I_s and I_d, are processed through a portrait animator to generate an augmented face. In the feature aggregation stage, both the original and augmented images are passed through a pre-trained face recongition model to extract and aggregate the face embeddings $\hat{y}$ for verification, yielding the aggregated feature $\hat{y}_{tta}$. Aggr. denotes aggregation.

the generated data, enhancing the overall robustness and accuracy. In particular, our method shows strong performance when there are variations in poses, demonstrating its robustness in handling non-frontal face images. In summary, our main contributions are as follows. We introduce Pose-TTA, which enhances the performance of pre-trained face recognition models during inference without requiring additional training. To improve robustness, we propose a weighted feature aggregation that mitigates biases and distortions in synthetic data generated by the portrait animator. Finally, we demonstrate the generalisation of Pose-TTA through extensive experiments across six training datasets and five model architectures.

II. PROPOSED METHOD

The overall framework of the proposed Pose-TTA is shown in Fig. 1. Our framework consists of two main stages: head pose alignment and feature aggregation. In the head pose alignment stage, the Face Selector takes two face images, I_1 and I_2, as inputs and determines the source image I_s, which retains the identity, and the driving image I_d, which provides the target pose. Then, I_s and I_d are passed through a pre-trained portrait animator to generate an augmented face $\hat{I}$. In the Feature Aggregation stage, both the original and multiple augmented images are processed through a pre-trained face recognition model to extract the face embedding $\hat{y}$. Finally, the aggregated feature $\hat{y}_{tta}$, a weighted combination of the augmented and original features, is used for verification.

A. Head Pose Alignment

Face selector. The Face Selector is responsible for selecting the most suitable face pose for transformation in order to minimise identity loss and distortion. A critical issue in face frontalisation is that synthesising a frontal view from a partially occluded face can lead to facial information distortion. The Face Selector takes two images as input and utilises an off-the-shelf head pose extractor [41] to estimate yaw, which is a key factor influencing facial information loss. Based on this, the image with the smaller absolute yaw value (p_2) is selected as the source image I_s, while the image with the larger absolute yaw value (p_1) is designated as the driving image I_d. Note that the off-the-shelf head pose estimator is not used during the pose augmentation process, thereby eliminating dependence on its performance.

Portrait animation. The goal of portrait animation is to transform the source image I_s into an image $\hat{I}$ that replicates the head pose of the driving image I_d. For this process, we employ LivePortrait [18] as the portrait animator. The key reasons for this choice are: (1) comparable performance to diffusion-based methods [19, 59, 63] while achieving faster inference speed; (2) independent control of facial expressions and head pose; and (3) unlike methods that require explicit roll/pitch/yaw inputs, LivePortrait conditions the transformation directly on a driving image, eliminating dependence on the head pose estimator's accuracy. However, directly inputting I_s and I_d into the portrait animator can lead to unexpected distortions when there is a substantial difference in head pose between the two images. This happens because the model must generate the occluded side of the face from I_s, which can introduce artifacts.

To mitigate this issue, we leverage facial symmetry. Using the yaw values extracted by the Face Selector, we identify cases where I_s and I_d have opposite facial directions (i.e., when the product of their yaw values is negative). In such cases, we horizontally flip I_s to roughly align its face direction with I_d, reducing the discrepancy between the two images. Furthermore, we experimentally observe that if the source image also mimics the facial expression of the driving image, it biases features such as eye shape, mouth shape, and facial contours toward those of the driving image. Therefore, we perform augmentation only on head pose. Finally, we obtain the augmented image $\hat{I}$ with minimal distortion from the pose transformation process.

B. Test Time Adaptation

If a candidate set of augmentations $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_{|\mathcal{T}|}\}$ is selected at test time, conventional test-time augmentation can be formulated as:

$$\hat{y}_{tta} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \Theta_{\text{target}}(\tau_i(x)), \tag{1}$$

TABLE I

EFFECTIVENESS OF POSE-TTA. TTA REFERS TO TEST-TIME AUGMENTATION. BASELINE RESULTS WITHOUT TTA USE THE AGGREGATION OF ORIGINAL AND FLIPPED EMBEDDINGS. ALL EXPERIMENTS USE $w_{real} = 0.75$. † INDICATES OUR TRAINED MODELS.

Method	Backbone	Train Data	TTA	Accuracy (%) CPLFW [64]	CFP-FP [43]	LFW [22]	CALFW [65]	AgeDB [37]	Avg
AdaFace [27]	ResNet101	WebFace12M [66]	✗	94.57	99.24	99.82	96.12	98.00	97.55
			✓	95.40 +0.83	99.23 -0.01	99.83 +0.01	96.08 -0.04	97.87 -0.03	**97.68**
		WebFace4M [66]	✗	94.63	99.27	99.83	96.05	97.90	97.54
			✓	95.28 +0.65	99.24 -0.03	99.80 -0.03	96.10 +0.05	98.02 +0.12	**97.69**
		MS1MV3 [12]	✗	93.92	99.09	99.83	96.02	98.18	97.41
			✓	94.75 +0.83	99.16 +0.07	99.83	96.20 +0.18	98.27 +0.09	**97.64**
		MS1MV2 [11]	✗	93.53	98.67	99.80	96.12	98.05	97.23
			✓	94.28 +0.75	98.67	99.82 +0.02	96.10 -0.02	98.05	**97.38**
		CASIA-Webface† [61]	✗	90.12	97.37	99.32	93.60	94.88	95.06
			✓	90.17 +0.05	97.49 +0.12	99.32	93.60	95.05 +0.17	**95.13**
		DCFace† [28]	✗	82.98	91.37	98.58	91.90	90.65	91.10
			✓	84.97 +1.99	92.10 +0.73	98.58	91.82 -0.08	90.78 +0.13	**91.65**
AdaFace [27]	ViT	WebFace4M [66]	✗	94.97	98.94	99.80	96.03	97.48	97.44
			✓	95.08 +0.09	99.10 +0.16	99.82 +0.02	96.03	97.22 -0.26	**97.45**
	ResNet50	WebFace4M [66]	✗	94.17	98.99	99.78	95.98	97.78	97.34
			✓	94.83 +0.66	98.89 -0.10	99.82 +0.04	95.95 -0.03	97.73 -0.05	**97.44**
		CASIA-Webface [61]	✗	90.02	97.04	99.37	93.43	94.40	94.85
			✓	90.35 +0.32	97.06 +0.02	99.30 -0.07	93.52 +0.09	94.17 -0.23	**94.88**
	ResNet18	WebFace4M [66]	✗	92.28	97.80	99.58	95.52	96.48	96.33
			✓	92.85 +0.57	97.91 +0.11	99.58	95.50 -0.02	96.35 -0.13	**96.44**
		CASIA-Webface [61]	✗	87.00	94.81	99.22	92.65	92.68	93.27
			✓	87.87 +0.87	95.34 +0.53	99.32 +0.10	92.97 +0.32	92.68	**93.64**
ArcFace [11]	ResNet101	WebFace4M [66]	✗	94.35	99.21	99.78	96.00	97.95	97.46
			✓	95.02 +0.67	99.24 +0.03	99.78	95.88 -0.12	98.03 +0.08	**97.59**
		CASIA-Webface† [61]	✗	89.98	97.23	99.47	93.52	94.63	94.97
			✓	90.08 +0.10	97.53 +0.30	99.38 -0.09	93.63 +0.11	94.32 -0.31	**94.99**
		DCFace† [28]	✗	82.50	90.49	98.68	92.05	90.83	90.91
			✓	84.22 +1.72	91.27 +0.78	98.65 -0.03	92.03 -0.02	90.70 -0.07	**91.37**

where Θ_{target} is a neural network trained on the target dataset, and x is an input image. In our case, Θ_{target} represents the pre-trained face recognition model. The set $\tau_i(x)$ consists of four representations when the Face Selector designates an image as the source: the original image, the augmented image, the horizontally flipped original image, and the flipped augmented image, incorporating facial symmetry. When an image is designated as the driving image by the Face Selector, $\tau_i(x)$ includes only two representations: the original image and the flipped original image.

Feature aggregation. [45] observed that simply averaging predictions from TTA-transformed images can sometimes degrade accuracy by turning correct predictions into incorrect ones. Based on this, we introduce a weighting mechanism into the TTA process, assigning lower weights to synthetic images due to potential distortions, thereby mitigating negative effects while preserving the benefits of TTA. The weighted feature aggregation is defined by the equation:

$$\hat{y}_{tta} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} w_{\mathcal{T}} * \Theta_{\text{target}}(\tau_i(x)). \tag{2}$$

Here, $w_{\mathcal{T}}$ is determined by the following conditions:

$$w_{\mathcal{T}} = \begin{cases} w_{syn}, & \text{if } \tau_i(x) \text{ is synthetic data} \\ w_{real}, & \text{if } \tau_i(x) \text{ is real data} \end{cases} \tag{3}$$

The values of w_{syn} and w_{real} are hyperparameters for balancing the embeddings, and in this paper, they are set to 0.25 and 0.75, respectively. Since the feature scale of $\hat{y}_{\text{tta}}$ can vary depending on the number of elements in $\tau_i(x)$, $\hat{y}_{\text{tta}}$ is ultimately normalised along the channel axis before face verification.

III. EXPERIMENTS

A. Experimental setup

Baselines. To validate the generalisation capability of the proposed Pose-TTA method, we select pre-trained models trained with a combination of various backbone models (ViT [14], ResNet18, ResNet50, ResNet101 [20]) and two training losses (AdaFace [27], ArcFace [11]). The datasets used in the experiments are as follows: Casia-WebFace [61], MS1MV2 [11], MS1MV3 [12], WebFace4M, WebFace12M [66], and the synthetic dataset DCFace [28]. These datasets contain 0.49M, 5.8M, 5.1M, 4.2M, 12M, and 0.5M facial images, respectively.

Dataset and evaluation. We evaluate the face recognition models on five datasets: CPLFW [64], CFP-FP [43], LFW [22], CALFW [65], and AgeDB [37]. CPLFW includes large pose variations, while CFP-FP consists of pairs of frontal and 90-degree side-profile faces. LFW contains variations in lighting and expression, with most images being frontal or near-frontal [1]. CALFW and AgeDB-30 feature images of the same person at different ages. We follow the evaluation protocol of CVLFace [26]

IV. RESULTS

A. Effectiveness of Pose-TTA

Table I presents the performance of face recognition models across multiple datasets, demonstrating the improvement achieved by applying our Pose-TTA during inference. The results show that applying TTA consistently improves performance on CPLFW and CFP-FP, with average accuracy gains of 0.72% and 0.19% across all models, respectively . Notably, these datasets contain significant head pose variations, suggesting that our augmentation strategy effectively

TABLE II

COMPARISON OF POSE AUGMENTATION METHODS USING A RESNET18 MODEL TRAINED ON CASIA-WEBFACE WITH ADAFACE.

Method	Accuracy (%)					
	CPLFW	CFP-FP	LFW	CALFW	AgeDB	Avg
Baseline	87.00	94.81	99.22	92.65	92.68	93.27
Frontalisation	86.83	94.41	99.23	92.65	92.65	93.15
Ours w/o flip	87.82	95.30	**99.32**	92.83	**92.72**	93.60
Ours	**87.87**	**95.34**	**99.32**	**92.97**	92.68	**93.64**

TABLE III

PERFORMANCE COMPARISON BASED ON WEIGHT IN WEIGHTED FEATURE AGGREGATION USING A RESNET18 MODEL TRAINED ON CASIA-WEBFACE WITH ADAFACE.

Aggr. Weight		Accuracy (%)					
w_{real}	w_{syn}	CPLFW	CFP-FP	LFW	CALFW	AgeDB	Avg
0.00	1.00	85.77	90.07	98.92	91.72	91.50	91.60
0.25	0.75	87.23	93.49	99.15	92.37	92.47	92.94
0.50	0.50	**87.88**	94.84	99.23	92.92	**92.75**	93.52
0.75	0.25	87.87	**95.34**	**99.32**	**92.97**	92.68	**93.64**
1.00	0.00	87.00	94.81	99.22	92.65	92.68	93.27

mitigates challenges from extreme poses and enhances model robustness in unconstrained face verification scenarios.

Conversely, for datasets dominated by frontal faces, such as LFW and CALFW, the performance change remains marginal. In some cases, we observe a slight drop in accuracy (0.04% on AgeDB-30), likely due to the introduction of unnecessary variations that do not contribute meaningful information to these datasets. These findings suggest that our method can enhance face verification performance in real-world scenarios where diverse head poses are encountered, while maintaining stable performance even when pose variations are minimal.

B. Comparison with Frontalisation.

As mentioned earlier, to quantitatively analyse the issue of facial distortion when using frontalisation in TTA, we compare our method with the approach presented in Table II. Notably, for datasets with diverse head poses (CPLFW, CFP-FP), frontalisation leads to a significant performance drop, resulting in an average degradation of 0.12% compared to the baseline model without TTA. In contrast, our method, which performs pose augmentation while minimising identity distortion, achieves an average performance improvement of 0.37%. Furthermore, when face direction alignment using flipping is omitted (denoted as 'Ours w/o flip'), by comparing the yaw angles of the two faces before feeding them into the portrait animator, we observe an average performance drop of 0.04%. This demonstrates that simple alignment using flipping, which leverages facial symmetry to compensate for large pose differences, effectively reduces identity distortion during the portrait animation process.

C. Effectiveness of Weighted Feature Aggregation.

Unlike traditional TTA methods, which apply transformations such as translation and rotation while preserving the original structure of objects within the image, our approach generates synthetic data by modifying the face within the image. Consequently, biases inherent in the generative model

Fig. 2. Head pose augmentation quality comparison. The proposed method minimises identity distortions by flipping the source image and applying portrait transformation.

and subtle distortions introduced during the generation process are unavoidable. To ensure reliable feature extraction when using both real and synthetic data, we propose a weighted aggregation method and present an ablation study on the weighting strategy in Table III. The results indicate that performance decreases when dependence on real data is reduced. Notably, when the weight assigned to real data is increased to 0.75, our method outperforms the conventional approach of averaging features from original and augmented data (i.e., w_{real} = 0.5). This demonstrates that our proposed method effectively mitigates the drawbacks of synthetic data, such as distortion and bias, ultimately contributing to improved performance.

D. Qualitative Comparison

We present the quality of head pose augmentation achieved by the proposed method in Fig. 2. As we pointed out, frontalisation methods inevitably generate unseen regions in the source image, leading to distortions in the original identity. Furthermore, when there is a significant difference in head pose between the source and driving faces, distortions occur during the portrait animation process (see the 'w/o flip' case). Consequently, by flipping the source image based on the yaw value obtained from the Face Selector and applying portrait transformation, we can minimise distortions from the original image and better preserve the identity.

V. CONCLUSION

In this paper, we present Pose-TTA, a novel approach that enhances pre-trained face recognition models during inference by utilising test-time augmentation with a portrait animator. Unlike traditional methods that rely on frontalisation, our approach aligns side-profile images, minimising identity distortion and improving verification accuracy. Additionally, our weighted feature aggregation strategy effectively addresses biases in synthetic data, ensuring a more reliable augmentation process. Extensive experiments across six datasets and five face recognition frameworks demonstrate the effectiveness and generalisation of Pose-TTA, showcasing its superiority over conventional TTA and face frontalisation methods. This approach offers an efficient way to improve face recognition models without extensive retraining and shows the potential to be applied to other face-related tasks in complex environments.

VI. ACKNOWLEDGMENTS

This work was supported by HDC Labs. We would like to thank Chansung Jung, Gunhee Lee and Chongmin Park for helpful discussions.

ETHICAL IMPACT STATEMENT

Potential Risks. We introduce Pose-TTA, a test-time augmentation method designed to enhance pose-invariant face recognition. While the method improves accuracy by proposing pose-agnostic TTA framework, it also raises ethical concerns related to fairness, privacy, and potential misuse.

One major risk lies in the inherent biases present in publicly available face recognition datasets [11, 12, 61, 66]. These datasets often exhibit demographic imbalances, leading to variations in recognition accuracy across different ethnicities, genders, and age groups [5, 31, 35, 48]. If these biases are not carefully addressed, the proposed method may cause performance disparities across different demographic groups. Additionally, large-scale face recognition datasets have raised serious ethical concerns, particularly regarding privacy violations and the lack of informed consent [3]. Many widely used datasets [11, 66] have been constructed by indiscriminately collecting web images, often without the subjects' approval. In some cases, the classification of "celebrities" has been broadly applied to individuals with an online presence, raising further ethical concerns. As a result of these issues, public access to several of these datasets has been revoked [54], emphasising the need for stricter data governance. These concerns underscore the necessity of ethical data collection practices and adherence to informed consent principles in face recognition research.

Moreover, recent advances in generative models across various domains—such as image and audio generation [24, 32, 51, 55]—have significantly expanded both the capabilities and the potential risks of AI-driven systems. In particular, the ability to synthetically manipulate head poses raises concerns about identity spoofing or potential misuse in unauthorised applications, such as deepfake generation [52]. Although our Pose-TTA does not explicitly involve identity synthesis, any system that modifies facial representations must consider risks related to authenticity and trustworthiness in biometric authentication.

Risk-Mitigation Strategies. To mitigate the aforementioned risks associated with dataset bias, privacy concerns, and potential misuse, we take several precautions in the design and implementation of Pose-TTA.

First, we analyse Pose-TTA's performance across multiple datasets [11, 12, 61, 66] and models to identify potential biases in recognition accuracy across demographic groups. This approach helps ensure that our augmentation technique does not disproportionately favour specific populations. Additionally, our method operates solely during inference as a test-time augmentation approach, without modifying the training process. This design minimises the risk of amplifying dataset biases, as the underlying recognition models remain unchanged. Furthermore, to address privacy concerns, we evaluated Pose-TTA using synthetic face datasets [28], thereby reducing reliance on sensitive real-world biometric data. This approach helps mitigate privacy risks related to data misuse and consent, while enabling robust model validation in privacy-preserving environments. Lastly, we promote the responsible use of Pose-TTA by restricting code access to organisations that agree to ethical usage guidelines. We emphasise its intended applications in research and authentication, while actively discouraging its use in surveillance, unauthorised identity manipulation, or adversarial AI applications.

Benefit-Risk Analysis. Despite the potential risks, Pose-TTA offers substantial benefits when applied responsibly. By addressing pose variations in face recognition, it improves model robustness without requiring additional training, making it an efficient and scalable solution for real-world applications. This improvement is particularly valuable for identity verification systems, where extreme pose variations often lead to recognition failures. Additionally, our test-time augmentation approach provides a computationally efficient alternative to traditional data augmentation methods, reducing the need for extensive retraining on pose-augmented datasets. Moreover, Pose-TTA maintains high performance even when applied to synthetic face datasets, underscoring its suitability for privacy-conscious applications. This demonstrates its ability to deliver accurate results while reducing dependence on sensitive personal data, thereby aligning with ethical AI principles and supporting privacy-preserving use cases. However, we acknowledge that face recognition technology remains ethically complex, and we strongly advocate for continued scrutiny, fairness evaluations, and regulatory oversight in its deployment. By ensuring transparency, responsible usage, and ongoing assessment of its societal impact, Pose-TTA can contribute positively to the advancement of ethical and reliable face recognition systems.

REFERENCES

[1] Z. An, W. Deng, Y. Zhong, Y. Huang, and X. Tao. Apa: Adaptive pose alignment for robust face recognition. In *Proc. CVPRW*, 2019. 1, 3

[2] A. Atzori, F. Boutros, N. Damer, G. Fenu, and M. Marras. If it's not enough, make it so: Reducing authentic data demand in face recognition through synthetic faces. In *Proc. FG*, 2024. 1

[3] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proc. WACV*, pages 3526–3535, 2023. 5

[4] S. Banerjee, J. Brogan, J. Krizaj, A. Bharati, B. R. Webster, V. Struc, P. J. Flynn, and W. J. Scheirer. To frontalize or not to frontalize: Do we really need elaborate pre-processing to improve face recognition? In *Proc. WACV*, 2018. 1

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 5

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proc. FG*, 2018. 1

[7] C. Chen and S. McCloskey. High-resolution image enumeration for low-resolution face recognition. In *Proc. FG*, 2024. 1

[8] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric*

Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13, 2018. 1

[9] X. Chu and T. Harada. Generalizable and animatable gaussian head avatar. In *Proc. NeurIPS*, 2024. 1

[10] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech*, 2018. 1

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. CVPR*, 2019. 1, 3, 5

[12] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *Proc. CVPRW*, 2019. 1, 3, 5

[13] Y. Deng, D. Wang, and B. Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *Proc. ECCV*, 2024. 1

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021. 3

[15] Q. Duan and L. Zhang. Boostgan for occlusive profile face frontalization and recognition. *arXiv preprint arXiv:1902.09782*, 2019. 1

[16] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He. High-fidelity face manipulation with extreme poses and expressions. *IEEE Transactions on Information Forensics and Security*, 16:2218–2231, 2021. 1

[17] Z. Gao, Q. Li, G. Wang, and L. Shen. Pointfaceformer: local and global attention based transformer for 3d point cloud face recognition. In *Proc. FG*, 2024. 1

[18] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 1, 2

[19] Y. Han, J. Zhu, K. He, X. Chen, Y. Ge, W. Li, X. Li, J. Zhang, C. Wang, and Y. Liu. Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *Proc. ECCV*, 2024. 2

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 3

[21] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun. Pose-guided photorealistic face rotation. In *Proc. CVPR*, 2018. 1

[22] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 3

[23] Y. Jang, K. Rho, J. Woo, H. Lee, J. Park, Y. Lim, B.-Y. Kim, and J. S. Chung. That's what i said: Fully-controllable talking face generation. In *Proc. ACM MM*, 2023. 1

[24] J. Jung, J. Ahn, C. Jung, T. D. Nguyen, Y. Jang, and J. S. Chung. Voicedit: Dual-condition diffusion transformer for environment-aware speech synthesis. In *Proc. ICASSP*, 2025. 5

[25] I. Kim, Y. Kim, and S. Kim. Learning loss for test-time augmentation. *Proc. NeurIPS*, 2020. 1

[26] M. Kim. CVLFace: High-Performance Face Recognition All-in-One Toolkit, 2024. 3

[27] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proc. CVPR*, 2022. 3

[28] M. Kim, F. Liu, A. Jain, and X. Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *Proc. CVPR*, 2023. 3, 5

[29] M. Kimura. Understanding test-time augmentation. In *Proc. NeurIPS*, 2021. 1

[30] D. Kwak, J. Jung, K. Nam, Y. Jang, J.-W. Jung, S. Watanabe, and J. S. Chung. Voxmm: Rich transcription of conversations in the wild. In *Proc. ICASSP*, 2024. 1

[31] D. Leslie. Understanding bias in facial recognition technologies. *arXiv preprint arXiv:2010.07023*, 2020. 5

[32] T. Li, Y. Tian, H. Li, M. Deng, and K. He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024. 5

[33] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proc. CVPR*, 2017. 1

[34] X. Luan, Z. Ding, L. Liu, W. Li, and X. Gao. A symmetrical siamese network framework with contrastive learning for pose-robust face recognition. *IEEE Trans. on Image Processing*, 2023. 1

[35] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021. 5

[36] Q. Meng, X. Xu, X. Wang, Y. Qian, Y. Qin, Z. Wang, C. Zhao, F. Zhou, and Z. Lei. Poseface: Pose-invariant features and pose-adaptive loss for face recognition. *arXiv preprint arXiv:2107.11721*, 2021. 1

[37] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proc. CVPR*, 2017. 3

[38] Z. Ou, L. Yao, T. Wu, and F. Liu. Aerialface: A light weight framework for unmanned aerial vehicle face recognition. In *Proc. FG*, 2024. 1

[39] E. Ozturk, M. Prabhushankar, and G. AlRegib. Intelligent multi-view test time augmentation. In *Intl. Conf. Image Proc.*, 2024. 1

[40] Y. Qian, W. Deng, and J. Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *Proc. CVPR*, 2019. 1

[41] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *Proc. CVPRW*, 2018. 2

[42] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, 2015. 1

[43] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *Proc. WACV*, 2016. 3

[44] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag. When and why test-time augmentation works. *arXiv preprint arXiv:2011.11156*, 1(3):4, 2020. 1

[45] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag. Better aggregation in test-time augmentation. In *Proc. CVPR*, 2021. 1, 3

[46] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. In *Proc. NeurIPS*, 2019. 1

[47] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov. Motion representations for articulated animation. In *Proc. NeurIPS*, 2021. 1

[48] T. Sixta, J. C. Jacques Junior, P. Buch-Cardona, E. Vazquez, and S. Escalera. Fairface challenge at eccv 2020: Analyzing bias in face recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 463–481. Springer, 2020. 5

[49] Y. Song and F. Wang. Qgface: Quality-guided joint training for mixed-quality face recognition. In *Proc. FG*, 2024. 1

[50] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proc. CVPR*, 2020. 1

[51] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 5

[52] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 5

[53] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proc. CVPR*, 2017. 1

[54] R. Van Noorden. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834):354–359, 2020. 5

[55] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023. 5

[56] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *Signal Processing Letters*, 25(7):926–930, 2018. 1

[57] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proc. CVPR*, 2018. 1

[58] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *Proc. ICLR*, 2022. 1

[59] H. Wei, Z. Yang, and Z. Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 2

[60] Y. Xie, H. Xu, G. Song, C. Wang, Y. Shi, and L. Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *Proc. ACM SIGGRAPH*, 2024. 1

[61] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1, 3, 5

[62] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu. Dual-attention gan for large-pose face frontalization. In *Proc. FG*, 2020. 1

[63] B. Zeng, X. Liu, S. Gao, B. Liu, H. Li, J. Liu, and B. Zhang. Face animation with an attribute-guided diffusion model. In *Proc. CVPR*, 2023. 2

[64] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7):5, 2018. 3

[65] T. Zheng, W. Deng, and J. Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. 3

[66] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proc. CVPR*, 2021. 1, 3, 5