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

Your Face, Your Privacy: Combating Unauthorized Usage

Atul Kumar, Akshay Agarwal, and Nalini Ratha
Trustworthy BiometraVision Lab, IISER Bhopal, India
University at Buffalo, USA
{atulk23, akagarwal}@iiserb.ac.in, nratha@buffalo.edu

2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3315-5341-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/FG61629.2025.11099478

***Abstract*— The high performance of current deep face recognition systems and their unauthorized usage have raised a severe concern for privacy in the physical, adversarial, and digital domains. To protect privacy, users are exploring several ways, and one such method that recently gained attention is individuals deliberately obscuring their faces with their hands, presumably to avoid facial recognition technology. Since deep face recognition algorithms can handle partial tampering of faces, this raises a critical question of whether these deliberate attempts can protect privacy. In the literature, no evaluation exists that showcases that this type of hiding can bypass the face recognition algorithms. Therefore, in this first-ever study, we have performed extensive research by first developing multiple nose and mouth occlusion datasets using synthetic patches and real-life objects. Our extensive experimentation reveals several interesting observations reflecting the fact that even when a patch is a face patch extracted from an unseen subject, it can fool the face recognition networks. Further, not only face recognition networks, but also it is observed that the proposed patches are effective in deceiving the soft biometric classifier, i.e., the classifier detecting the gender and ethnicity of individuals.**

I. INTRODUCTION

The growing use of facial recognition technology has raised concerns about individuals' privacy, forcing them to adopt various techniques to hide their identity, such as covering their faces with hands and scarves or strategically positioning themselves to avoid being recognized. Fig. 1 demonstrates this behavior using various examples from the MAFA [6] dataset. In addition, a recent article released in January 2024 by the New York Post titled "What is 'nose cover' — and why are Gen Zers doing it in family photos?" delves into the contemporary phenomenon where younger individuals intentionally conceal specific facial features, particularly the nose and mouth, as a means of protecting their identity[1]. However, a prominent question arises **"to what extent do these types of occlusions help in preserving the privacy of an individual or obscure their facial recognition?"**, especially when the current deep face recognition algorithms yield high accuracy under the availability of partial faces [11], [12]. The closest work to this work is based on the generation of face images where the faces are partially occluded through masks [31] or using large devices [25]. Wang et al. [31] have developed the masked face dataset, and Qiu et al. [25] have showcased the use of large objects such as mobile phones and big stickers to evaluate the vulnerability of face recognition algorithms. The primary limitations of these works are that they do not utilize natural objects and hence do not reflect the physical world setting or occlude a large portion of the face, such as face masks.

Fig. 1: Examples showcasing the mediums humans can use intentionally or unintentionally to hide their identity by fooling the recognition algorithms.

To overcome these limitations and effectively address privacy concerns inspired by current trends of simple face occlusion, we propose multiple nose/mouth occlusion datasets by simple patches based on ethnicity and skin-agnostic tone. *The patches can also be seen as blind obfuscation in terms of no access to the deep face recognition networks*. In other words, the patches that hide the mouth and nose features do not require knowledge of deep face recognition algorithms. We utilize the 105-classes-pin dataset[2], which contains a collection of celebrity images, to study the effect of partial face feature occlusion on deep face recognition. To investigate the influence of occlusions, various forms of patches are generated to adequately understand the role of patch features in deceiving the recognition algorithms.

The current existing studies, which are similar to our work, only address the concern of face recognition; however, we are aware that face modality is rich in containing other essential attributes such as gender and ethnicity. This research expands its scope by investigating the impact of the proposed patches in identifying these soft biometrics attributes, with a particular focus on identifying gender and ethnicity. For that, we have utilized the benchmark datasets, namely UTKFace [34] and Fairface [14], that are balanced in terms of gender and ethnicity. In brief, the contributions of this research are:

- Simple partial face feature tampering datasets are proposed, and the vulnerability of several deep face recognition networks is analyzed;
- An extensive experimental study is performed to understand the sensitivity of partial face tampering in

[1]https://nypost.com/2024/01/11/lifestyle/what-is-nose-cover-and-why-are-gen-z-teens-doing-it-in-family-photos/

[2]https://www.kaggle.com/datasets/hereisburak/pins-face-recognition

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

identifying soft biometric attributes.

- An inpainting approach without a reference image mimics real-world scenarios where the original face is unavailable, offering insights into how well modern algorithms restore facial regions across different identities.

II. Related Work

Deep face recognition that utilizes large amounts of data has shown tremendous success and surpasses the human level performance [23], [28], [29]. It is observed that the networks utilizing large-scale face images are generally acquired from online sources without consent. Further, it is shown that selfies or group images uploaded on personal social media platforms have faces that have not consented to be used as datasets for training models. The above issues show the valid case of personal privacy and security of an individual's digital platform. To tackle this, users and researchers are exploring methods for privacy-enhancing recognition algorithms [27]. The methods can be broadly divided into two groups: (i) utilizing noises to perturb the face images [33], [7] and (ii) use of external components such as masks and devices [1], [25], [31]. For example, Chhabra et al. [3] and Mirjalili and Ross [22] proposed a method utilizing adversarial perturbations to conceal distinct facial characteristics to achieve gender and smile anonymization. In contrast to image-based perturbation, Xue et al. [32] proposed the adversarial perturbations of the feature space to conceal the identities of facial images. Besides utilizing the perturbation to anonymize the soft biometric attribute, several adversarial attacks and morphing algorithms are proposed to fool the face recognition algorithms. For example, Goswami et al. [8], [9] have proposed several black-box methods to fool deep face recognition algorithms. Komkov et al. [16] attach a basic paper sticker to a hat, while Frearson et al. [5] showcase using visible light to trick the systems. In addition, Zhu et al. [35] presented a makeup-based attack to effectively bypass recognition models. In contrast, Majumdar et al. [19] investigated the effectiveness of partial morphing of certain facial regions of deepface recognition networks. While these methods are found effective in either anonymizing the soft biometric attributes or fooling the face recognition algorithms, they have several drawbacks: (i) need access to models to learn perturbation and are computationally heavy, (ii) modify facial features drastically [8]. Apart from the perturbation-based strategy, another formal case where synthetic data is used to train the model, but is it really as effective as real data [21].

The second school of thought studies the vulnerability of deep face recognition utilizing external components such as mobile phones, masks, and stickers. While face masks and external components can hide the identity of individuals, they drastically conceal their facial features. The proposed research acts as an intermediary between these two schools of thought, which do not utilize any learnable noise through the use of face recognition models or drastically hide the facial features. Further, the proposed occlusions are relevant to protect the privacy of identities and anonymize the soft attributes, including gender and ethnicity.

TABLE I: Overview of statistical characteristics of different datasets used in this paper. G and P represent the gallery and probe, respectively. M and F represent the male and female gender attributes, respectively, whereas W, B, A, and I represent the ethnicity attributes, namely white, black, Asian, and Indian, respectively.

Datasets	Task							
	Recognition		Gender and Ethnicity/Race Classification					
	G	P	M	F	W	B	A	I
105-classes-pin	105	315	-	-	-	-	-	-
UTKFace [34]	-	-	12,390	11,314	10,077	4522	3,432	3,975
FairFace [14]	-	-	51,778	45,920	18,606	13,789	26,043	13,835

III. Experimental Setup

In this section, we first describe the patches used to generate a face patch dataset exhibiting the current trend of occluding nose and mouth features to enhance privacy. The proposed dataset is not bound to the trend but aims to reflect the importance of occluded regions and characteristics of patches. Later, deep face recognition algorithms used to perform the recognition are described, followed by the networks used for soft attribute prediction.

A. Proposed Patch Face Datasets

To assess the performance of the occlusion on face recognition, we employed an unconstrained and challenging dataset, namely a 105-class Pinterest dataset comprising 105 subjects belonging to distinct celebrities. The images in the dataset are collected from Pinterest. Further, to analyze whether the proposed patch generation techniques effectively anonymize the soft biometric attributes, we have utilized two benchmark datasets: UTKFace [34] and FairFace [14]. The UTKFace [34] dataset provides a large-scale collection of "wild" 23,704 facial images in pose, expression, illumination, and variations resolution. Each subject has a single image with annotations for age, gender (11,314 females, 12,390 males), and ethnicity (White, Black, Asian, Indian, Others). FairFace [14] offers a balanced dataset of 97,698 images with gender (45,920 females, 51,778 males) and seven racial categories (White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, Latino). Table I provides an overview of the datasets mentioned.

We employed various face patch simulations on a designated dataset to systematically evaluate the influence of occlusions on facial recognition performance and soft attribute prediction. Our primary patch shape form is rectangular and focuses on 40% of the face width. However, for generalization, we have also created 25% and 50% of the face width of rectangular shapes, and an oval-shaped patch is also introduced with 40% of the face width to analyze the results on the shape and size of the patch. Fig. 2 (leftmost block) visualizes the diverse subjects and patches of datasets used in this study. Fig. 2 (middle) shows the example image of different shapes and sizes of the patches. Also, some real-life object patches, such as hand and mobile patches, have a comparison with the ethnicity and skin-tone patch to the

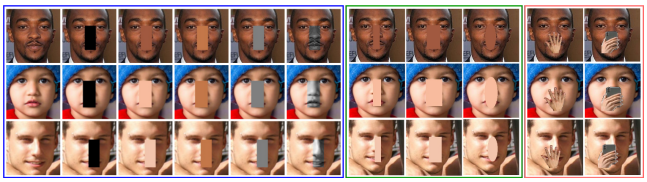

Fig. 2: A few images of 105-classes-pin (1^{st} row), UTKFace (2^{nd} row), and FairFace (3^{rd} row) dataset. From left to right, images of clean and five different patches (leftmost is BP, followed by ChP, FPP, GrP, and GrSP) showcase the variation they bring in the visual appearance of faces. The leftmost blue block shows uniform rectangular patches, the middle green block contains images of varying patch sizes and types (oval), and the rightmost red block contains images with real-life patches.

real-life object. Fig. 2 (rightmost block) shows the example image of the real-life object patch consisting of hand and phone patches. A detailed description of the spectrum of patches introduced in this research is as follows:

1) ***Black patch (BP)***: This uniform black patch simulates scenarios where objects obstruct the face, and the complete detail of the face feature is missing.
2) ***Cheek patch (ChP)***: While perturbing facial regions, we have utilized a personalized feature approach. A patch is extracted from each subject's cheek region, replicating self-occlusion by mimicking their unique skin tone and texture. We assert that this kind of perturbation can also help us understand if the facial features are present and duplicated; such a perturbation can fool the networks.
3) ***Face-pixel patch (FPP)***: In this patch attack, in place of utilizing the face patches from each perturbing subject, we utilize a global patch template. This global patch template is outsourced from an unseen subject outside the original dataset used for evaluation. In contrast to the above ChP perturbation, this injects identity-independent facial features. The FPP and ChP patches are close to real-world methods, where skin pixels (hand) are used for privacy preservation.
4) ***Gray patch (GrP)***: A neutral gray patch is employed to analyze the effect of non-descript occlusions with a medium level of contrast.
5) ***Gray-scale patch (GrSP)***: This patch converts the images to grayscale, enabling the assessment of how the loss of color information impacts recognition performance.
6) ***Hand Patch (HP)***: In this perturbation, a patch simulating a hand is applied over the facial region, replicating a common real-world scenario where a person covers their face partially with their hand.
7) ***Phone Patch (PhP)***: The phone patch is designed to simulate the scenario where a person is holding a phone to their face, obscuring key facial features such as the nose, mouth, and part of the cheek.

Each dataset comprises eight probe/testing sets, including one set of *clean images (C)* and seven different patches.

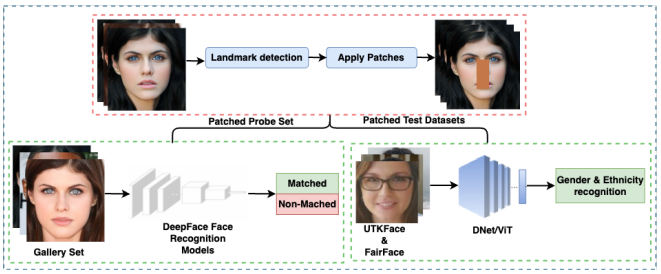

Fig. 3: Schematic diagram of the proposed face, gender, and ethnicity recognition setup.

These patches include five synthetic rectangular patches: *black patch (BP)*, *cheek patch (ChP)* in which the cheek pixel of each individual is taken to create a patch, *gray patch (GrP)*, *gray-scale patch (GrSP)*, and *face-pixel patch (FPP)*, as well as two real-life object patches: *hand patch (HP)* and *phone patch (PhP)*. Fig. 3 illustrates the architecture diagram of the face, gender, and ethnicity setup, highlighting the flow from patch generation, training, and testing on patched datasets. The pseudo-code for the generation of this patched dataset is mentioned in the supplementary material.

B. Deep Face Recognition Networks

To comprehensively understand the impact of proposed patches, we have used state-of-the-art (SOTA) deep face recognition models: (i) ArcFace [4], (ii) Dlib [15], (iii) VGG-Face [24], (iv) FaceNet [26], (v) FaceNet-512, (vi) DeepFace, and (vii) SphereFace [18]. To perform matching using the feature extraction from the above pretrained models, we have employed distance matching metrics, including cosine (Cos) [30], Euclidean (Euc) [20], and EuclideanL2 (EucL2) [20]. For face recognition, we have utilized an image as a gallery, and three images of each subject are used as a probe set. The patch images are generated by modifying the probe set and used for evaluation only. By integrating these sophisticated models, distance metrics, and a methodical comparison framework, our methodology offers a nuanced and comprehensive exploration of face recognition technology.

C. Occluded Soft Biometric Recognition

In recent years, soft biometrics has emerged as a powerful tool to enhance recognition systems by helping segregate the ample search space. Further, as mentioned, faces inherently contain soft biometric attributes, including gender and ethnicity. The protection of not only identity but also these attributes is equally essential. Henceforth, we have conducted soft attribute anonymization through black-box patches for the first time. For the prediction of gender and ethnicity soft attributes, we have used two current and SOTA models –ViT-B/32 (ViT) [10], and DenseNet-121 (DNet) [13]. The recent Vision Transformer (VT) network utilizes a self-attention mechanism, whereas DenseNet-121 is a conventional pure convolutional neural network (CNN) architecture with residual connections for adequate gradient flow. To comprehensively evaluate the effectiveness of

TABLE II: Face recognition accuracy reflecting the privacy gained using different patches occluding nose and mouth features. The values on clean face images are bold, the values of the best-performing patch are underlined and blue-colored, and the values of the second-best-performing patch are colored green for better visibility and understanding. The lower the value on the patch images, the better the privacy.

CNN	Distance Metrics																	
	Cosine (Cos)						Euclidean (Euc)						Euclidean-L2 (EucL2)					
	C	BP	ChP	GrP	GrSP	FPP	C	BP	ChP	GrP	GrSP	FPP	C	BP	ChP	GrP	GrSP	FPP
ArcFace	**94.92**	21.59	71.11	72.38	92.38	74.60	**70.48**	12.38	42.86	42.54	68.25	45.71	**93.65**	14.29	63.17	64.13	91.75	66.98
Dlib	**90.79**	23.81	59.68	67.62	86.03	63.81	**94.60**	41.27	77.78	81.59	89.84	78.73	**93.02**	37.46	74.60	78.73	88.89	77.46
VGG-Face	**88.25**	35.87	51.75	55.56	85.40	59.05	**89.21**	37.14	53.97	57.14	86.03	61.27	**89.21**	37.14	53.97	57.14	86.03	61.27
SFace	**81.27**	10.16	42.86	49.21	71.43	46.03	**85.40**	12.06	50.79	59.68	77.78	53.02	**74.60**	6.35	33.33	38.73	64.13	35.56
Facenet	**80.32**	11.11	27.94	30.16	75.87	31.75	**67.94**	7.94	19.68	19.68	63.81	20.32	**64.76**	3.17	9.21	12.70	59.68	12.06
Facenet512	**52.70**	1.59	4.13	4.76	49.52	6.03	**91.43**	53.33	58.10	58.41	91.75	61.27	**94.92**	59.05	65.08	66.98	91.75	72.06
DeepFace	**45.40**	1.27	17.78	23.79	44.13	25.71	**48.57**	4.76	21.27	29.52	45.71	30.48	**36.81**	0.32	10.79	14.29	34.29	15.56

the proposed anonymization approach, gender and ethnicity classification experiments are performed in the same and cross-dataset settings. The UTKFace [34] and FairFace [14] datasets are divided into training and testing. The patch subsets are generated on the test set of each dataset and are only used for evaluation. In other words, a clean training subset of both datasets individually is used for training. For training the gender classification networks, 12799 and 52757 images from UTKFace and FairFace are used, and the testing has been performed on 538 and 600 images from both datasets, respectively. Similarly, training of ethnicity classification has been done using 13204 and 43365 images and testing using 1022 and 1200 images of $UTKFace$ and $FairFace$, respectively. The last few layers of each model are fine-tuned using the batch size of 32, a learning rate of 0.0001, and an Adam [2] optimizer.

IV. Experimental Results and Analysis

In this section, we first present face recognition results and analysis, explaining whether such partial occlusion can preserve individuals' privacy. Later, a comprehensive understanding of the anonymization of soft biometric attributes is presented. Finally, an inpainting approach to mitigate the effects of patches to protect from intruders to fool the system.

A. Occluded Face Recognition

Our research involved an in-depth examination of how seven cutting-edge facial recognition models, three different distance metrics, and eight probe sets of different shapes and sizes interacted to preserve privacy. Table II shows the comprehensive analysis of face recognition on various models, along with the combination of 3 distance metrics on all the different probe sets containing the patches of 40% face width. The analysis can be broadly divided into three categories: (i) effectiveness of deep face recognition (DFR) models, (ii) impact of distance metric in matching, and (iii) privacy obtained through each patch used to occlude face features. It can be seen from the results that the ArcFace [4] and FaceNet512 yield the highest clean accuracy across all the models used for face recognition. For example, the ArcFace (with cosine) and FaceNet512 (with EuclL2) models obtained a clean face recognition accuracy of 94.92%. It is interesting to note that even the images are clean due to their visual complexity since they are acquired from an unconstrained environment, which can lead to low recognition accuracy. For example, the difference between the best-performing network, ArcFace, and the worst-performing network, DeepFace, is 49.52%, where the cosine distance metric is used for the match. It shows that the correct use of the network is also vital for their massive difference in efficiency. Apart from the chosen face recognition network, a distance-matching function is also crucial for comparing gallery and probe features. For instance, the ArcFace and FaceNet512, which yield the highest performance, are found sensitive concerning the distance metric. The ArcFace model performs best with cosine (Cos), while FaceNet512 shows the highest performance with EuclideanL2 (EucL2).

Fig. 4: Examples reflecting the success (top row) and failure (bottom row) of ArcFace + Cosine (first six columns) and Dlib + Euclidean (last six columns) on clean and in the presence of nose/mouth occlude patches. Images from left to right represent clean images followed by the occlusion using BP, ChP, FPP, GrP, and GrSP patches, respectively.

In terms of patches used, the black patch shows the highest level of privacy and drastically degrades each network's accuracy. The prime reason for its success may be that it completely occludes the face features and ensures they are missing for matching. However, it is worth noting that, in our case, only a tiny fraction of face information is occluded, which shows that the network is susceptible to such small perturbations. It is interesting to note that other patches that utilize the skin pixel (cheek pixels of the same person or different person) and grayscale provide sufficient privacy. For instance, while the black patch (BP) reduces the performance of ArcFace + Cos from 94.92% to 21.59%, the ChP (coming from the subject-specific cheek region) also reduces the performance by 23.81%. The performance gap increases further to 30.48% when the ArcFace model is used with the EucL2 distance metric, and the ChP patch is used to protect the privacy of individuals. In place of using the patch coming from the same individual, when subject-agnostic patch (FPP) pixels are used for attack, they are found comparable or better in degrading the performance of each network. The Dlib + Euc shows an accuracy reduction to 77.78% and 78.73% from 94.60% when an attack is performed using subject-specific and subject-agnostic patches, respectively.

TABLE III: Face recognition accuracy reflecting the privacy gained through the use of different patches occluding nose and mouth features on **varied shapes and sizes of the patch**. The values on clean face images are bold, the values of the best-performing patch are underlined and blue-colored, and the values of the second-best-performing patch are colored green for better visibility and understanding.

Model	C	BP	ChP	GrP	GrSP	FPP
Oval Patch						
ArcFace	**94.92**	26.98	72.20	74.44	92.33	76.19
Rectangular Patch size: 25%						
ArcFace	**94.92**	50.16	82.54	83.49	91.11	82.22
Rectangular Patch size: 50%						
ArcFace	**94.92**	16.19	62.22	62.22	91.43	56.19

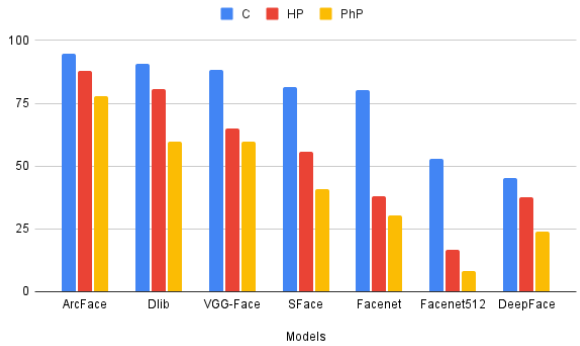

Fig. 5: Face Recognition accuracy on both real-world patches, i.e., HP and PhP, on best-performing cosine distance metrics.

It is interesting to observe that not only color patches (ChP or FPP) but also gray patches can protect privacy and significantly reduce the performance of each network. For example, the color ChP patch reduces the performance VGG-Face + Cos from 88.25% to 51.75% as compared to 55.56% obtained from the gray patch (GrP). Fig. 4 shows the correctly and incorrectly recognized images in the presence of different patches used. It shows that the proposed patches can provide privacy to each gender (no gender bias) and work on frontal and pose-inherited faces.

Now, moving towards different shapes and sizes of the patch, along with the effect of real-life object occlusion. Regarding the oval patch, the accuracy values are almost similar to the rectangular patch, as it covers nearly identical regions, but only due to the curved nature, it exposes some of the facial landmarks and features, increasing by up to 1% to 2% in accuracy. Table III shows the result analysis of the best-performing ArcFace (with cosine) combination on different shapes and sizes of the patch. When the patch covers 50% of the face width, the ArcFace-cosine accuracy drops from 21.59% (at 40% width) to 16.19%. Conversely, reducing the patch to just 25% of the face width raises accuracy to 28.57%, since exposing key landmarks around the nose and mouth helps in improving the performance. We have also applied some real-life object patches with an accuracy of 87.94% and 77.78% on hand and mobile patches, respectively, on the ArcFace model with cosine distance. Fig. 5 shows the bar graph accuracy visualization for the face recognition on the best-performing cosine distance metrics. Refer to the supplementary material for detailed results of the patch's different shapes and sizes, along with the real-life object patch.

Fig. 6: Few correctly classified and misclassified images when used for **gender** classification across patches. The first row shows the correctly classified samples of UTKFace and FairFace, respectively, and the second row shows the incorrectly classified samples of UTKFace and FairFace, respectively.

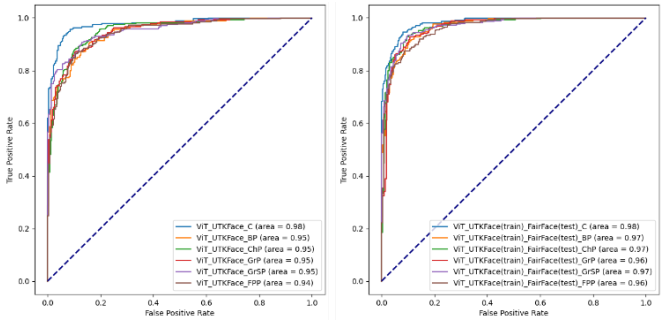

Fig. 7: ROC of gender prediction of ViT-B/32 model on seen (left) and unseen (right) datasets.

B. Occluded Soft Biometric Recognition

As can be seen from the above analysis, the proposed patches can provide varied levels of privacy, but back the theory that people recently used to protect themselves by covering their noses (in turn occluding their mouths as well). To further strengthen the findings of this research, we have performed extensive experiments to find whether such phenomena can also anonymize their soft biometric attributes, i.e., gender and ethnicity/race. First, we analyze the impact of patch-based occlusion on gender classification. In the end, an analysis concerning ethnicity/race classification is provided to show how effective patches are in hiding soft attributes. For gender and ethnicity classification, we have used two state-of-the-art deep networks: ViT-B/32 (based on attention mechanism) and DenseNet-121 (a pure convolutional neural network).

1) Occluded Gender Classification: The evaluation of ViT-B/32 [10] and DenseNet-121 [13] (DNet-121) models for occluded gender recognition across UTKFace and FairFace datasets is shown in Table IV. Similar to the analysis concerning face recognition, gender classification accuracy can be described in terms of the following factors: (i) the capacity of the network, (ii) the anonymization impact of each patch, and (iii) the robustness and generalizability of the patches.

TABLE IV: Gender prediction accuracy under seen and unseen dataset settings in the presence of different patches used to occlude the nose and mouth facial regions. While it is observed that the patches are found less effective in fooling transformer models, they are found significantly effective in fooling convolutional networks.

Train Dataset	Model	Test Dataset											
		UTKFace						FairFace					
		C	BP	ChP	GrP	GrSP	FPP	C	BP	ChP	GrP	GrSP	FPP
UTKFace	ViT-B/32	**93.49**	87.73	87.92	87.55	88.66	86.78	**93.00**	90.86	88.79	91.03	91.90	89.66
	DNet-121	**90.86**	79.93	71.56	64.13	75.09	72.22	**86.83**	75.86	68.10	61.38	74.14	70.86
FairFace	ViT-B/32	**91.08**	88.48	86.62	88.48	86.06	86.78	**96.67**	93.62	94.31	94.66	93.45	91.72
	DNet-121	**86.80**	79.55	74.91	75.65	74.35	73.75	**91.33**	85.69	83.79	82.76	81.55	80.34

TABLE V: Gender prediction accuracy under seen and unseen dataset settings in the presence of different size and shape patches (oval patches and rectangle patches of 25% & 50%). While it is observed that the patches are found less effective in fooling the ViT-B/32 model, they are found significantly effective in fooling convolutional networks.

Model	Train Dataset	Test Dataset											
		UTKFace						FairFace					
		C	BP	ChP	GrP	GrSP	FPP	C	BP	ChP	GrP	GrSP	FPP
ViT-B/32	UTKFace	**93.49**	84.48	88.12	86.02	86.4	86.78	**93.00**	88.45	88.79	90.52	89.66	88.62
	FairFace	**91.08**	85.63	87.74	87.93	88.70	87.93	**96.67**	92.93	94.31	94.48	94.48	94.31
ViT-B/32	UTKFace	**93.49**	89.46	90.8	90.61	88.31	89.08	**93.00**	93.1	91.72	92.59	91.38	90.34
	FairFace	**91.08**	87.36	89.27	88.89	88.31	89.08	**96.67**	94.48	93.62	94.48	95.00	94.31
ViT-B/32	UTKFace	**93.49**	85.63	86.02	86.21	87.93	86.4	**93.00**	89.48	89.83	90.52	90.17	88.28
	FairFace	**91.08**	86.4	88.12	88.12	87.55	86.78	**96.67**	94.41	92.93	93.28	93.62	92.24

TABLE VI: Ethnicity prediction accuracy under seen and unseen dataset settings in the presence of different patches used to occlude the nose and mouth facial regions. While it is observed that the patches are found to be less effective in fooling convolutional networks, they are found to be significantly effective in fooling transformer models.

Train Dataset	Model	Test Dataset											
		UTKFace						FairFace					
		C	BP	ChP	GrP	GrSP	FPP	C	BP	ChP	GrP	GrSP	FPP
UTKFace	ViT-B/32	**88.06**	60.30	64.12	61.41	72.96	54.97	**69.42**	57.78	61.36	60.25	65.68	55.19
	DNet-121	**93.84**	82.21	87.74	86.53	90.05	87.94	**64.58**	60.62	62.96	61.73	67.90	62.96
FairFace	ViT-B/32	**78.25**	67.41	71.36	69.63	72.84	69.26	**77.59**	59.80	66.63	60.50	67.04	62.41
	DNet-121	**87.42**	75.68	79.75	78.89	84.44	78.77	**81.31**	64.22	65.23	64.42	76.84	63.12

Fig. 8: A few correctly classified and misclassified **ethnicity** images when perturbed using the proposed patches. The first row shows the correctly classified samples of UTKFace and FairFace, respectively, and the second row shows the incorrectly classified samples of UTKFace and FairFace, respectively.

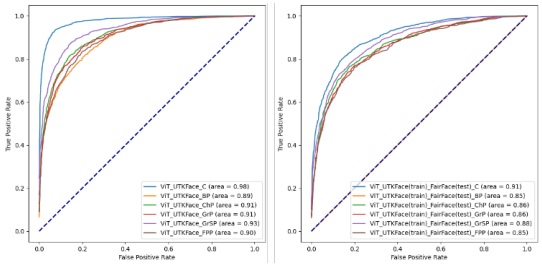

Fig. 9: ROC of ethnicity prediction of ViT-B/32 model on seen (left) and unseen (right) datasets.

It is observed that the ViT model is found to be highly effective as compared to the DenseNet model in classifying gender on clean images. When the ViT-B/32 is trained on clean images of UTKFace and tested on the clean images of UTKFace, it yields an accuracy of 93.49% as compared to

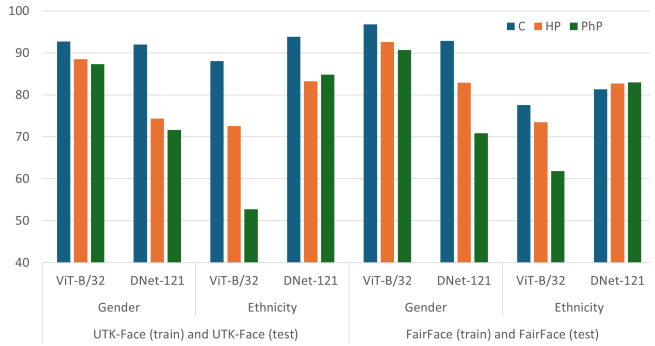

Fig. 10: Gender and ethnicity prediction accuracy under seen dataset training and testing setting under real-life patches bar graph visualization. In the case of ViT, a phone patch is found to be more effective than a hand patch; however, the DenseNet model, the majority of the time, has higher sensitivity to hand patches.

90.86% obtained using the DenseNet-121 (DNet-121) model. The ViT model is found robust in performing gender classification even if the dataset images are not seen at the time of training. As can be seen from the results, the ViT model yields slightly lower performance (0.49%) when an unseen dataset is used for evaluation in comparison to the same training-testing dataset evaluation. Interestingly, the black

TABLE VII: Ethnicity prediction accuracy under seen and unseen dataset settings in the presence of different size and shape patches (Oval patches and rectangle patches of 25% & 50%). While it is observed that the patches are found less effective in fooling the D-Net-121 model, they are found significantly effective in fooling transformer models.

Model	Train	Test											
		UTKFace						FairFace					
		C	BP	ChP	GrP	GrSP	FPP	C	BP	ChP	GrP	GrSP	FPP
DNet-121	UTKFace	**93.84**	82.91	88.24	87.94	91.76	88.54	**64.58**	63.33	66.54	66.05	69.01	65.68
	FairFace	**87.42**	65.63	70.45	70.45	76.48	68.54	**81.31**	76.17	83.58	81.73	86.42	81.23
DNet-121	UTKFace	**93.84**	87.04	90.45	89.65	90.85	90.35	**64.58**	65.56	66.91	66.42	67.65	65.58
	FairFace	**87.42**	67.14	69.85	70.75	74.87	69.25	**81.31**	79.14	82.47	82.59	84.32	81.36
DNet-121	UTKFace	**93.84**	77.09	84.82	83.32	90.15	86.33	**64.58**	57.53	61.11	59.63	66.91	60.25
	FairFace	**87.42**	62.81	64.42	63.62	77.49	62.41	**81.31**	74.32	78.77	77.16	85.19	76.42

patch that drastically reduces face recognition performance is found to be less effective in fooling gender classification networks, especially ViT. Further, the network's vulnerability decreases when it is evaluated on an unseen dataset. In other words, the relative difference between the clean images and patch images' gender classification accuracy is reduced further if the network is trained on an unseen dataset. For instance, the relative difference when the ViT model is trained on UTKFace and tested on UTKFace is 5.73%, which reduces to 2.14% when evaluated on an unseen FairFace dataset. Surprisingly, in contrast to the ViT model, the pure CNN architecture is found highly vulnerable to patch attacks. Further, this vulnerability is not affected even when the model is tested on an unseen dataset. When the DNet-121 model is trained on clean UTKFace and tested on the black patch images of the UTKFace and FairFace datasets, it shows a drop of 10.93% and 10.97%, respectively. Surprisingly, the grayscale patch, found less effective in providing privacy, is found significantly effective (even higher than the black patch) in anonymizing the gender and is found resilient in training and testing conditions. Grayscale patches (GrP) and subject-agnostic (FPP) patches are found to be effective in hiding gender information. As can be seen in the majority of the cases when the FairFace dataset is used for training, the face pixel patch (FPP) yields the highest reduction in gender classification accuracy compared to other patches. The correctly classified and misclassified occluded gender prediction of each patch is shown in Fig. 6. The ROC curves of gender classification are shown in Fig. 7. We have also performed experiments on varied shapes and sizes of the patch along with the real-life object patch on soft biometrics recognition, and in gender, we achieve only a 1% to 2% accuracy difference. Table V shows the best model performing analysis. The detailed experimental results are mentioned in the supplementary material.

2) Occluded Ethnicity Classification: Similar to gender classification, ethnicity classification is also performed under seen and unseen dataset settings using ViT-B/32 and DenseNet-121 (DNet-121). As shown in Table VI, in contrast to the gender classification, for ethnicity classification, the DNet-121 model is found to be highly effective as compared to ViT. The performance of DNet-121 for ethnicity classification on the UTKFace train test is 5.78% higher than ViT on the clean images of the dataset. The DNet-121 model shows consistent effectiveness across different datasets, including unseen dataset settings, except when trained on UTKFace and evaluated on FairFace. Further, the ViT model is found vulnerable to patches when utilized for ethnicity classification compared to the DNet-121 model, which shows higher robustness when used for ethnicity classification. The analysis shows that there is no silver bullet (single network) that is robust in handling patches when deployed for different tasks of gender and ethnicity classification. Similar to face recognition, for ethnicity classification, the stealthy rate of black patches is found to be high compared to other patches most of the time. However, the face pixel patch is also able to reduce the performance significantly and is found to be effective in fooling ViT. For instance, the ViT model, which yields 88.06% ethnicity classification performance on UTKFace, suffers a drop to 54.97% when FPP is applied to the testing images. We believe such an extensive analysis is missing in the literature, highlighting that ViT effectively performs tasks accurately, but its resiliency against different perturbations drastically varies across tasks. The correctly classified and misclassified occluded ethnicity prediction of each patch is shown in Fig. 8. The macro average ROC curve for predicting ethnicity using the best-performing model ViT-B/32 on both the seen and unseen datasets is shown in Fig. 9. Similar to gender recognition, the accuracy difference is only 1% to 2% in ethnicity recognition when experiments are performed on varied shapes and sizes of the patch, along with the real-life object patch. Fig. 10 shows the performance of gender and ethnicity prediction on HP and PhP. Table VII shows the best-performing model analysis; refer to the supplementary material for detailed experimentation results.

C. Effect of Inpainting in Mitigating Partial Occlusion

The results demonstrate that the applied patches successfully deceive deep face recognition models, but they also present a potential vulnerability by enabling intruders to exploit the system. To address this issue, we employ a state-of-the-art inpainting technique, the Mask-Aware Transformer (MAT) [17], to restore the occluded regions and counteract the effects of these patches. Specifically, we use two strong synthetic patches found in our analysis, i.e., black and cheek, along with real-life object patches to evaluate the impact of inpainting on face recognition performance. Table VIII presents the accuracy of the face recognition models after applying inpainting. It is observed that while inpainting alleviates the occlusion effect, its success is not uniform across different scenarios. For instance, the black patch, which is highly effective in fooling face recognition models, has its accuracy significantly improved when inpainted using the MAT algorithm. However, when the inpainting method attempts to reconstruct facial regions occluded by real-life objects, such as hands and phones, the performance of face recognition models degrades. For example, ArcFace + Cosine achieves an accuracy of 87.94% with a hand-occluded patch, but this drops to 73.97% after the MAT algorithm is applied to reconstruct the region. This analysis highlights that while inpainting effectively reverses the effects of artificial patches, it fails to address occlusions caused by real-life objects.

TABLE VIII: Face recognition accuracy on the face images after inpainting. The bold values show the accuracy of the clean set. The green values show the best-performing inpainting.

Model	Distance Metrics														
	Cosine					Euclidean					Euclidean-L2				
	Clean	BP_inp	ChP_inp	HP_inp	PhP_inp	Clean	BP_inp	ChP_inp	HP_inp	PhP_inp	Clean	BP_inp	ChP_inp	HP_inp	PhP_inp
ArcFace	**94.92**	72.70	75.24	73.97	69.84	**70.48**	26.98	28.57	33.33	32.70	**93.65**	65.40	66.03	66.03	58.73
Dlib	**90.79**	72.70	74.92	76.51	64.13	**94.60**	84.44	85.71	89.21	78.10	**93.02**	83.81	83.81	87.94	77.78
VGG-Face	**88.25**	56.19	56.19	53.97	52.06	**89.21**	57.14	58.73	55.56	52.38	**89.21**	57.14	58.73	55.56	52.38
SFace	**81.27**	50.79	57.78	42.86	35.24	**85.40**	53.65	55.87	54.29	45.40	**74.60**	40.63	45.40	31.75	27.30
FaceNet	**80.32**	27.30	34.29	24.13	22.22	**67.94**	14.92	18.41	12.06	10.79	**64.76**	12.38	14.29	9.52	7.30
FaceNet-512	**52.70**	8.57	9.52	6.35	5.40	**91.43**	53.65	54.92	52.70	49.52	**94.92**	57.78	60.00	61.27	59.68
DeepFace	**45.40**	32.70	33.33	33.97	23.17	**48.57**	37.14	41.27	35.87	28.25	**36.81**	21.90	24.76	22.22	13.65

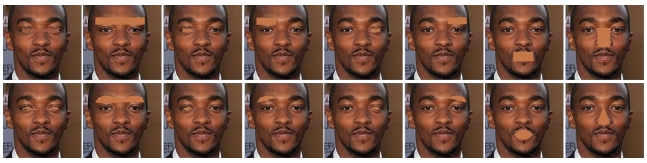

Fig. 11: Visualization of different facial attributes patches obscuring different attributes of the faces. 1^{st} row shows the rectangular patches and 2^{nd} row shows the facial attribute-shaped patches, specifically extracted from the cheek region.

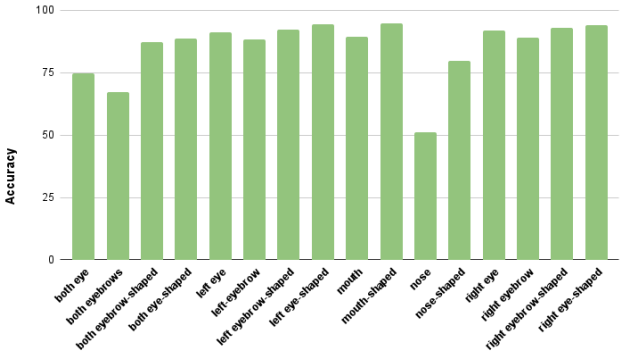

Fig. 12: Face recognition accuracy reflecting the privacy gained by using different patches occluding different facial attribute features.

Although these occlusions help preserve individual privacy, their potential to provide intruders with undue advantages, particularly in cases of real-life object occlusion, presents a critical challenge that we believe can be mitigated through strong inpainting methods.

V. Ablation Study

To validate our observation that occluding the nose and mouth regions offers the most effective privacy preservation, we conducted an ablation study by creating individual attribute patches targeting specific facial features: eyes, eyebrows, nose, and mouth. Specifically, we generated patches for the following facial attributes: both eyes, both eyebrows, both eyebrow-shaped, both eye-shaped, left eye, left eyebrow, left eyebrow-shaped, left eye-shaped, mouth, mouth-shaped, nose, nose-shaped, right eye, right eyebrow, right eyebrow-shaped, and right eye-shaped. Fig. 11 visualizes these patches, while Fig. 12 presents the face recognition accuracy using the ArcFace model combined with the Cosine distance metric for each attribute patch applied to the best privacy-preserving black patch. The results reveal that occluding the nose region (*nose* and *nose-shaped*) significantly reduces face recognition accuracy from **94.92%** (clean) to 51.11% and 79.87%, respectively, underscoring the nose's critical role in recognition systems. Interestingly, patches targeting the eyebrows also substantially degrade recognition performance, with the *Both Eyebrows* patch reducing accuracy to 67.41%. Notably, eyebrow occlusions are more effective in preserving privacy compared to eye occlusions, where patches such as *both eyes*, *both eye-shaped*, *left eye*, and *right eye-shaped* resulted in only minimal decreases in accuracy (ranging from 74.92% to 93.92%). This indicates that eyebrows carry more significant biometric information than eyes in face recognition models. Also, mouth patches showed a moderate impact, with the standard *mouth* patch reducing accuracy to 89.45%, while the *mouth-shaped* patch maintained a high accuracy of 94.88%. These findings confirm that the nose and eyebrows are more influential in face recognition models, and their occlusion can more effectively preserve privacy. This ablation study supports our primary approach of occluding the nose and mouth regions to enhance privacy protection against deep face recognition systems. Refer to the supplementary material for detailed results of these facial attribute patches on all other patches.

VI. Conclusion and Future Work

The rising need to protect privacy or avoid social media trolls due to unauthorized access to facial images requires significant attention, especially with the success of deep face recognition. Recently, people have explored hiding their noses and mouths using their hands to preserve privacy. Our extensive experiments with the novel occluded datasets reveal that specific patches can significantly degrade the performance of face recognition networks and soft biometric attribute classifiers when these regions are protected. Hence, we assert that these occluded images can be used for social media sharing to avoid their unconsented use for AI model attribute extraction. Further, we demonstrated that strong inpainting can be a viable solution to prevent any malicious use of these patches. In the future, we aim to advance privacy by developing invisible patches so that social media sharing of images is not hampered and privacy concerns are resolved.

Acknowledgement

A. Agarwal is partially supported through the ECRG grant of ANRF. A. Kumar is supported through the JRF fellowship of UGC, India.

ETHICAL STATEMENT

Author/Reviewer Checklist:

1) Did you read the Ethical Impact Statement Guidelines document (provided above)? **Yes**
2) Is it clear that all studies and procedures described in the paper were approved (or exempted) by a valid ethical review board? Alternatively, is a valid and sufficient justification provided for why the oversight of an ethical review board was not required? **Yes**
3) Does the ethical impact statement provide a clear, complete, and balanced discussion of the potential risks of individual harm and negative societal impacts associated with the research? Note that this includes harm to research participants as well as harm to other individuals that may be affected by use, misuse, or misunderstanding of the research. **Yes**
4) Does the ethical impact statement describe reasonable, valid, and sufficient use of risk-mitigation strategies by the authors to lessen these potential risks? Alternatively, if relevant strategies were not used, is a valid and sufficient justification for this provided? **Yes**
5) Does the ethical impact statement provide a valid and sufficient justification for how/why the potential risks of the research are outweighed by the risk-mitigation strategies and potential benefits of the research? Note that papers with serious potential risks that are not outweighed by risk-mitigation strategies and potential benefits may be rejected. **Yes**
6) If the paper involves human subjects, are all of the following sub-boxes checked?
    a) Does the main paper describe whether/how informed consent and/or assent were obtained from participants? If consent and/or assent were fully or partially obtained, were the methods used to do so valid? If not fully obtained, does the ethical impact statement provide a valid and sufficient justification for this? **Not Applicable**
    b) Does the main paper state whether the participants explicitly consented to the use of their data in the manner described in the paper? For example, if the data was or will be shared with third parties, does it state that the participants explicitly agreed to this sharing? If some uses were not explicitly consented to, does the ethical impact statement provide a valid and sufficient justification for this? **Not Applicable**
    c) Does the main paper explain whether/how participants were compensated? If participants were compensated, does the ethical impact statement provide a valid and sufficient justification for the form and amount of compensation provided? **Not Applicable**
    d) If the research involves any special or vulnerable populations (e.g., minors, elderly individuals, prisoners, refugees and migrants, individuals with disabilities, individuals with mental illness, or patients in medical settings), does the ethical impact statement provide a valid and sufficient explanation of how the rights, well-being, and autonomy of such individuals were safeguarded in the research? **Not Applicable**

REFERENCES

[1] A. Agarwal, N. Ratha, M. Vatsa, and R. Singh. When sketch face recognition meets mask obfuscation: Database and benchmark. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–5, 2021.

[2] S. Bock and M. Weiß. A proof of local convergence for the adam optimizer. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

[3] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta. Anonymizing k-facial attributes via adversarial perturbations. *arXiv preprint arXiv:1805.09380*, 2018.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[5] M. Frearson and K. Nguyen. Adversarial attack on facial recognition using visible light. *arXiv preprint arXiv:2011.12680*, 2020.

[6] S. Ge, J. Li, Q. Ye, and Z. Luo. Detecting masked faces in the wild with lle-cnns. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2682–2690, 2017.

[7] A. Goel, A. Singh, A. Agarwal, M. Vatsa, and R. Singh. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2018.

[8] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *International Journal of Computer Vision*, 127:719–742, 2019.

[9] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[10] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2022.

[11] L. He, H. Li, Q. Zhang, and Z. Sun. Dynamic feature learning for partial face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7054–7063, 2018.

[12] B. Huang, Z. Wang, G. Wang, K. Jiang, Z. Han, T. Lu, and C. Liang. Plface: Progressive learning for face recognition with mask bias. *Pattern Recognition*, 135:109142, 2023.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[14] K. Kärkkäinen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[15] D. E. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[16] S. Komkov and A. Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *IEEE International Conference on pattern recognition*, pages 819–826, 2021.

[17] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia. Mat: Mask-aware transformer for large hole image inpainting. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[18] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[19] P. Majumdar, A. Agarwal, R. Singh, and M. Vatsa. Evading face recognition via partial tampering of faces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[20] M. Malkauthekar. Analysis of euclidean distance and manhattan distance measure in face recognition. In *Third International Conference on Computational Intelligence and Information Technology (CIIT 2013)*, pages 503–507. IET, 2013.

[21] R. Manju, A. Kumar, and A. Agarwal. On which data distribution (synthetic or real) we should rely for soft biometric classification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6238–6247. IEEE, 2025.

[22] V. Mirjalili and A. Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 564–573, 2017.
[23] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi. Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1642–1646, 2007.
[24] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*. British Machine Vision Association, 2015.
[25] H. Qiu, D. Gong, Z. Li, W. Liu, and D. Tao. End2end occluded face recognition by masking corrupted features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6939–6952, 2022.
[26] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.
[27] R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 13583–13589, 2020.
[28] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.
[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.
[30] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.
[31] Z. Wang, B. Huang, G. Wang, P. Yi, and K. Jiang. Masked face recognition dataset and application. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023.
[32] H. Xue, B. Liu, X. Yuan, M. Ding, and T. Zhu. Face image de-identification by feature space adversarial perturbation. *Concurrency and Computation: Practice and Experience*, 35(5):e7554, 2023.
[33] X. Yang, D. Yang, Y. Dong, W. Yu, H. Su, and J. Zhu. Delving into the adversarial robustness on face recognition. *arXiv preprint arXiv:2007.04118*, 2, 2020.
[34] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017.
[35] Z.-A. Zhu, Y.-Z. Lu, and C.-K. Chiang. Generating adversarial examples by makeup attacks on face recognition. In *IEEE International Conference on Image Processing*, pages 2516–2520, 2019.