My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization

Umur A. Çiftçi
Binghamton University
uciftci@binghamton.edu

Gokturk Yuksek
Binghamton University
gokturk@binghamton.edu

İlke Demir
Intel Labs
ilke.demir@intel.com

Abstract

Recently, productization of face recognition and identification algorithms have become the most controversial topic about ethical AI. As new policies around digital identities are formed [22], we introduce three face access models in a hypothetical social network, where the user has the power to only appear in photos they approve. Our approach eclipses current tagging systems and replaces unapproved faces with quantitatively dissimilar deepfakes. In addition, we propose new metrics specific for this task, where the deepfake is generated at random with a guaranteed dissimilarity. We explain access models based on strictness of the data flow, and discuss impact of each model on privacy, usability, and performance. We evaluate our system on Facial Descriptor Dataset [61] as the real dataset, and two synthetic datasets with random and equal class distributions. Running seven SOTA face recognizers on our results, MFMC reduces the average accuracy by 61%. Lastly, we extensively analyze similarity metrics, deepfake generators, and datasets in structural, visual, and generative spaces; supporting the design choices and verifying the quality.

1. Introduction

Face recognition and identification have been one of the most interesting research topics in computer vision [71]. Although the research has contributed to the collective knowledge; the applications have been controversial because of maleficent motivations, deployment of immature products, and bias in data and algorithms. Furthermore, consequences of these faulty products mostly remained unpunished due to the lack of cyber laws around digital identity. Early Google products tagging some skin tones as non-human [3], Meta being forced to delete 2 billion face embeddings [1], or ClearView AI aiming to encode every human's face [2] are just the tip of the iceberg when it comes to facial identity preservation and the impacts of lack thereof.

As a defense mechanism, counter manipulation techniques such as blurring [64], masking [57], and noise addition [66] encounter traditional face recognition. Similarly, adversarial generation and confiscation methods [15, 72] are developed for tricking deep learning based face detection and recognition systems. Although effective, these approaches disable face recognition by altering the image content, thus, it is a matter of time that face recognition algorithms are trained and armed against adversarial attacks [70, 35]. Joining these two fronts, our approach can be thought as a based-on-need masking mechanism that does not break the image continuity, and misleads face recognition systems with fake faces, using deepfakes for good.

We would like to demonstrate this privacy enhancing deepfake scenario on a new form of social network. In current social platforms, access rights are defined per image, which friends (or connections) are allowed to see. However, we all appear in hundreds of photos voluntarily or involuntarily, we believe that the access rights should be designed per face, where everyone has freedom over which photos they appear. In current systems, this is poorly handled with tagging/untagging choices, however the photo, and the faces, live on the platform forever even if all faces are untagged. The basic principle behind ”My Face My Choice” (MFMC) is that, your face is replaced with a dissimilar enough deepfake in the view of those you do not grant access. We quantitatively analyze and verify that created deepfakes (1) are not similar to the original face by the embedding distance, (2) are not similar to any other face by using synthetic source images, (3) approximate the original age and gender in the image by the selected distance metric, and (4) preserve the original head-pose and expression by the selected generator. Depending on three access levels providing contextual integrity, we propose solutions with different restrictions where the face embedding may only be stored on the client. Our contributions are listed as:

- A novel privacy enhancing anonymization system with quantitatively dissimilar deepfakes,
- Quantitative analysis of deepfakes in image, structure, embedding, reconstruction, and generative spaces, and

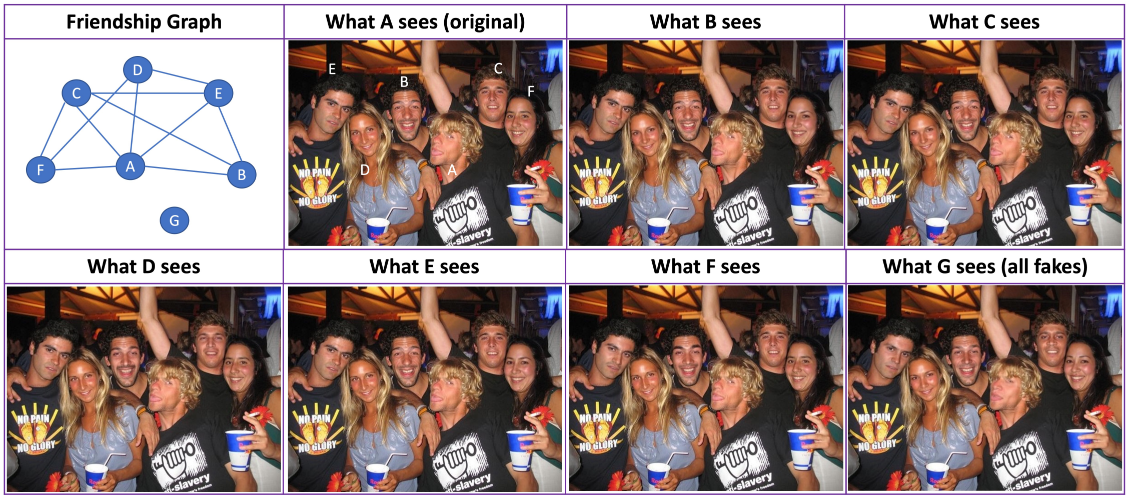


Figure 1. Our anonymization system masks faces with quantitatively dissimilar deepfakes in social photos, according to the friend graph and access rights. Here, A uploads a photo, B-F sees versions where non-friend faces are faked, and G sees everyone as fakes.

- Design and analysis of the system based on face access rights on a social photo sharing network.

MFMC can create deepfake versions of photos with more than 20 people, based on complex access rights given by people in the photo. In addition, we sketch these face access priorities in three levels to enable isolation of the face embedding versus detection and elimination of the face embedding. We use a diverse social image database [61] for the real dataset, and we evaluate MFMC using two fake datasets (i) 10,000 deepfakes created by StyleGAN [29] with random distribution, and (ii) 10,000 deepfakes created by Generated.Photos [27] with equal distribution across skin tones and genders. We analyze our system using four different GANs, with five distance metrics, and against seven face recognition approaches. We believe that MFMC is the first approach to extensively utilize face embeddings to create useful deepfakes as an adversary to face recognition.

2. Previous Work

2.1. Face Recognition and Identification

As a structured domain with lots of data, faces have been the most interesting playground for detection, recognition, and identification algorithms since Viola and Jones [63]. Early systems, deploying those algorithms without proper generalization and accuracy guarantees, caused universal consequences as mentioned in the introduction.

Using a set of known faces from their training process, current state-of-the-art face identification algorithms [39, 5, 34, 65, 37] try to find the identity of the test sample by calculating its deep feature representation. While training, every face identification network is trained with an appropriate loss function such as angular loss [37] or softmax [65], that minimizes deep feature distances between faces of the same person and maximizes it between faces with different IDs. During inference, test sample's deep feature representation is computed and used in a one-to-many similarity comparison with known faces to determine its ID. If the distance is smaller than a threshold, face identification network will predict both faces to belong to the same person.

Parallel to deep learning increasing the power of such face recognition systems [60, 67, 37, 39, 5, 34], novel open-source systems [6, 60, 53, 14] and proprietary software [4, 42] emerge for face detection and identification. As these systems become more popular, privacy implications [47, 26] and impacts on critical populations [8, 19, 74] of such have also been analyzed closely.

2.2. Face Generation and Deepfakes

Generative Adversarial Networks (GANs) [20] has been the stepping stone for creating realistic human face images that are difficult to visually discern from real faces [44]. Since then, the realism and resolution of the images significantly increased due to changing the generator from getting input latent code only, to the beginning of the network as an input, and then using mapping networks that transform the latent code to transmit into multiple layers of the network [28]. In this process, style concept emerged to control these layers with adaptive instance normalization [16, 25]. Other recent trends in face generation include providing additional noise maps to the generator to increase variability [55, 29]. There has also been some GANs that aim to create privacy-preserving faces or face masks [40, 26, 24].

Recent advances in face generation allowed easier and realistic generation of deepfakes and other facial manipulations. These manipulations and deepfakes can be grouped in four main categories as (1) novel face synthesis [28, 29] including responsible generation [13], (2) identity swap [33, 38, 46, 9, 45], (3) attribute manipulation [25], and (4) expression modification [62]. Our system utilizes the results of (1) as the source for non-existing faces to create deepfakes and the GANs in (2) for creating deepfakes with the same expressions, poses, and attributes.

2.3. Privacy Preserving Faces

Overall, face biometrics is one of the most personally identifiable information that is released without proper access rights [41]. As an adversary to face recognition, face obfuscation methods such as blocking and blurring [64], noise addition [66], and inpainting [57] have been proposed. However, such approaches break the continuity of the image and reveal that the image is obviously altered. In contrast, de-identification and anonymization approaches keep the image intact and modify the face to trick the recognition systems. Such methods are proposed on images [43] and videos [18]; using model-based [21], GAN-based [68], or hybrid [58] approaches; for face detection [7], action recognition [51], and annotation [56]. Our approach is also a privacy-preserving system, however it works in a multi-person setting based on access rights and social graphs, bridging the gap between privacy-preserving algorithms and real-world platforms.

2.4. Adversarial Attacks

Another perspective in breaking face identification algorithms, especially for deep learning based methods, is adversarial attacks. Similar to the approaches in the previous section, adversarial attacks also change the image content, mostly in the reconstruction and generative spaces, instead of directly swapping the face. These attacks may be image perturbations [48], poisoning attacks on training data [73], or cloaking images right after capture [54]. Algorithmic manipulations and basic adversarial attacks are shorter term solutions for the ever-improving GANs, so we choose the path of "creating as many fake faces as possible" to explode the embedding space of face recognition approaches.

3. Privacy Enhancing Deepfakes

The main motivation of MFMC is to keep the face images and face embeddings as local as possible, using the social graph and the access rights set by the users. In a nutshell, when an image is uploaded from the client, the *friends* (i.e., users connected to the uploader) are optionally tagged. Remaining faces in the photo are replaced with deepfakes and the image is sent to the server with this metadata. The tagged friends see others based on their friendship and outsiders see everyone as fakes. In this section, we describe the deepfake target selection metrics, deepfake creation process, and the design of face access rights.

3.1. Target Face Query

Anonymization of a face through deepfakes requires replacing the source face with another face that fits the same facial frame in a realistic way. The new face needs to be different enough for the anonymization to be successful. In order to define this difference, we need a comparison metric which yields similar enough faces for contextual and visual continuity, and dissimilar enough to confuse face recognition systems. As the image space is prone to misalignment, illumination, and noise, we proceed with the face embedding space, where we also preserve main face attributes such as age and gender. In our framework, we use ArcFace [14] to extract face embeddings with 512 features. ArcFace, using a novel loss function, optimizes feature embeddings to enforce higher similarity for intra-class samples and diversity for the inter-class samples. Moreover, we utilize InsightFace [23] to perform gender and age based classifications to store the similarity in the latent space. Note that this classification is neither exposed, nor used explicitly in our system, it is only stored as a direction in the latent space to define the similarity bounds. Experiments in other metric spaces are documented in Sec. 4.1.3.

Having computed face embeddings, we need a metric to compare two embeddings in order to query the "best" synthetic face as a replacement for the original face. We experiment with (1) minimizing the embedding distance for finding the closest face, (2) maximizing the embedding distance for finding the furthest face, (3) minimizing the embedding distance along the age and gender directions in the latent space as for finding the closest face with similar general attributes, (4) maximizing the embedding distance along the age and gender directions in the latent space for finding the furthest face with similar general attributes, (5) randomizing (2) within a plausible threshold to diversify the created samples, and (6) randomizing (4) within a plausible threshold to diversify the created samples. These choices are further evaluated in Sec. 4. Based on the motivation of "creating as many fakes as possible" the randomization is needed and the dissimilarity threshold ensures that the deepfake is still not recognizable as the source face. Overall, the set of target faces are created as,

$$\begin{gathered} T(S) = \{I_i \in I\} \;\; where \\ ||E(I_i) - E(S)|| > \\ \max(||E(I_j) - E(S)||) - \sigma(||E(I_j) - E(S)||) \\ \forall j \in I \end{gathered} \tag{1}$$

where $I = I_0, I_1, \ldots, I_N$ represents all images in the synthetic dataset, S is the real source image, and $T(S)$ is the target image set. $||.||$ denotes ℓ_2 distance, $E(.)$ is the face embedding, and σ is the standard deviation.

3.2. Deepfake Generation

In our framework, anonymization of a face requires a deepfake technique that fulfills transferring the identity of a synthetic target face into to the original source face while preserving extrinsic attributes such as expression, head pose, gaze direction, and lighting. In addition, the deepfake generator should be able to apply this process to multiple faces independent of occlusions and interactions.

The traditional deepfake creation with a common encoder and identity-specific decoders is not generalizable for this task as training an identity specific decoder requires multiple images with various poses. Multiple face requirement multiplies the need for compute resources and needs storing face decoders. As storing faces and identities conflicts with our main motivation, we cannot use traditional deepfakes. Overall, the deepfake generator (i) should be generalizable, working with any face (both as a source and a target) without any prior training, (ii) should change the identity of the source with the target significantly, (iii) should preserve facial and environmental attributes such as facial expression, head pose, gaze direction, lighting consistency, and (iv) should not create visible artifacts.

To satisfy these requirements, we integrated multiple deepfake generators into our framework. For 3D face model based approaches, Nirkin et al. [46] morph a 3D face model created from a source image into target face's expression values and use Poisson Blending to merge the two models. FTGAN [38] applies Few-Shot Unsupervised Image-to-Image Translation [36] and combines it with SPADE [49] module in order to inject semantic priors for face-swapping. FSGAN [45] uses a two-stage network architecture. While the first stage network performs the expression transfer with the reenactment process, second stage blends the result of first stage network into target image using face inpainting. Lastly, we integrated SimSwap [9], which is a modification of the traditional deepfake method with an ID injection module for generalization to arbitrary faces. It also uses weak feature matching loss to realistically reflect the facial attributes while still preserving the source identity information. We evaluated these methods (Sec. 4.1.1) and concluded that SimSwap provides the best results on different datasets, measured by five different losses.

3.3. Face Access Models

The ability to generate deepfakes must be accompanied by a capable access model to strike a proper balance between the desire to participate in social media and to preserve individual privacy. Furthermore, it needs to accommodate varying degrees of privacy needs, instead of imposing a one-size-fits-all solution. We assume that social platform operators are willing participants in MFMC with incentives to improve user privacy and abide by their privacy guidelines. MFMC supports the following access models:

3.3.1 Disclosure by Proxy

In this model, all the faces in a picture except the ones chosen by the individual that uploads it to the social media platform are deepfakes. This is disclosure by proxy because anybody who is in the picture and wishes to reveal their real face must be allowed by the original submitter. Once the picture with real and deepfake faces are finalized, the client disposes of the original picture.

The privacy of individuals depend on the submitter respecting others' consent, or lack thereof, to have their real faces revealed. Since none of the real faces remain on the platform, this model extends the privacy protection of MFMC against information leakage where the social media platform is compromised. In addition, no extra persistent storage is necessary on the client for face embeddings, or on the server as only the deepfake version of a photo is kept.

3.3.2 Disclosure by Explicit Authorization

This access model incorporates the tagging feature available in most social media platforms, in which tagging establishes a link between the face and a user. Typically, the platform then requests the consent of the tagged person to be associated with the face. We build on this feature such that when the users consents, deepfake is not created. Likewise, when a tag is removed, the platform replaces the real face with a deepfake. All non-tagged faces are deepfakes by default.

Since the platform strictly requires individual consent for a face to be tagged, accidental disclosure by a proxy is not possible unlike in the previous access model. Individuals that do not have accounts on the social media platform, but have faces captured in a picture with or without consent, remain anonymous as their faces are always protected by a deepfake. Given the fluid nature of tag/untag operations, the platform must keep a permanent copy of all the real faces both to generate deepfakes for and to restore as necessary.

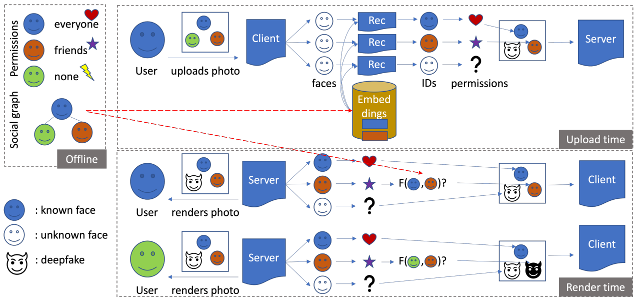


Figure 2. **Workflow for a three user network.** Green does not allow being seen by anybody, so its face embeddings are never shared, and it is always deepfaked on client. Blue grants to be seen by everyone, so it is always real. Orange only wants to be recognized by friends, so its face embedding is only shared with friends, and is deepfaked by the server at render time, per viewer.

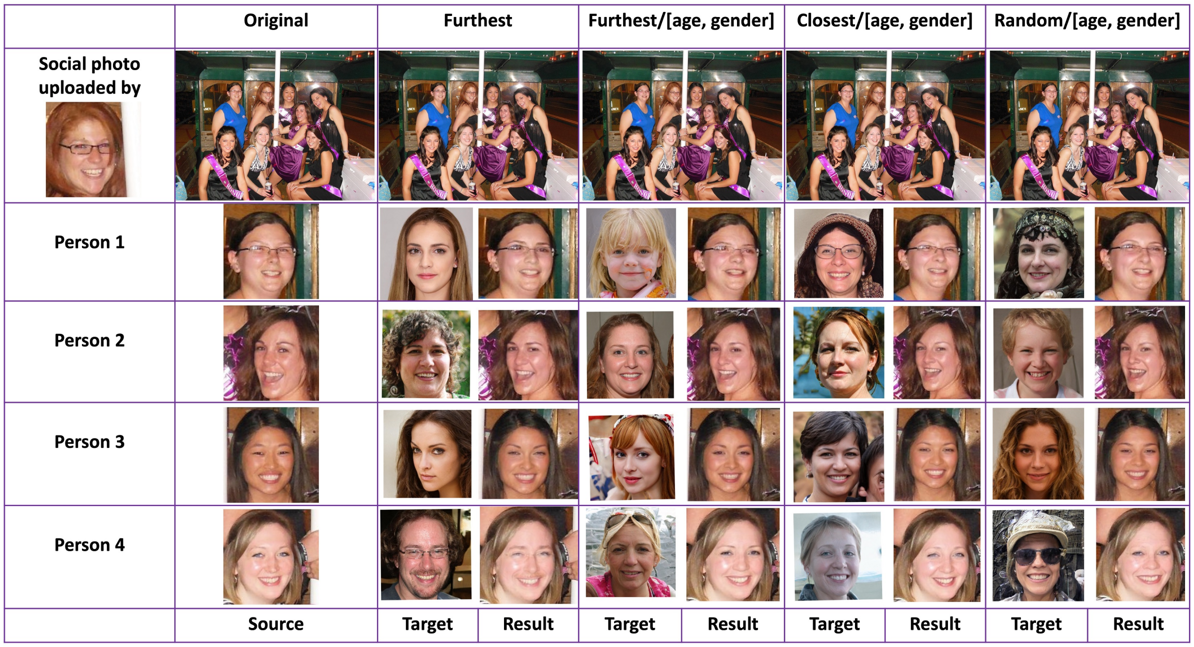


Figure 3. **MFMC Results per Similarity Query.** Source images and corresponding deepfakes created from closest/furthest/random target queries, and restricted by age and gender groups, are demonstrated.

Every tag/untag operation triggers a regeneration of the picture, thus, the performance requirements are higher compared to the previous model. The storage demands are also increased due to the need for persistently storing the original picture as well as the deepfake versions.

3.3.3 Access Rule Based Disclosure

Similarly to the previous model, all unknown faces are deepfakes by default during upload. Face embeddings of permitted friends/followers are kept on device, and if a face embedding is not on the device at upload time, it is unknown. Unlike the previous model, at render time, the platform consults a set of access rules supplied by face owners, to decide whether their face should be revealed (Fig. 2).

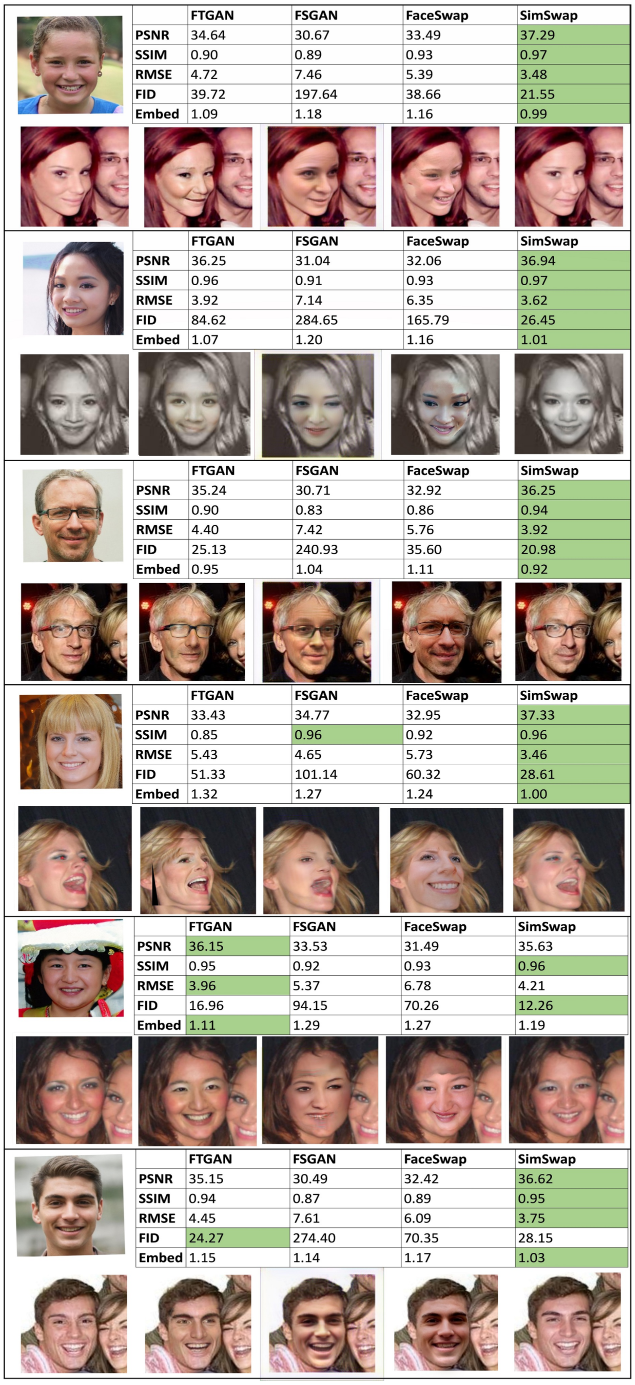
In their simplest form, access rules would allow a user to specify how their face should be revealed to friends or followers versus the general public. More sophisticated access rules could enable finer-grain control, allow different deepfakes for different observers, and possibly more.

The dynamic nature of different viewers possibly observing different versions of the same picture comes at the cost of increased storage and computation demands. More complex access rules demand more storage and computation.

4. Results

MFMC is implemented using InsightFace [23] library for face detection (which employs a pre-trained version of arcface [14]) and SimSwap [9] library for deepfake generation. We empirically selected InsightFace as a fast and accurate detector over OpenFace [6] and FaceNet [53], however the face detector can be swapped if the speed vs. accuracy trade off changes. Inference and training modules run on an NVIDIA GTX 1060 GPU. For the real dataset covering many social interactions in crowded environments, we utilize "party" subset of Facial Descriptors Dataset [61]. The dataset includes many configurations of 1282 people in 193 images, with various resolutions, poses, and illumination. In order to create the face manifold for MFMC to choose source images from, we use two datasets with unknown and uniform distributions. The first one contains 10,000 fakes created with StyleGAN [29], and the second one contains 10,000 fakes with equal skin tone, gender, and age distributions created by Generated.photos [27]. The similarity metric is face embedding distance unless otherwise is noted.

Fig. 3 contains MFMC results for an uploaded photo. For each face, we show the chosen target image from StyleGAN dataset and the created deepfake, according to four metrics as explained in Sec. 3.1. As seen in the furthest target image of Person 4, opposite genders may be chosen if there is no age and gender normalization. Similarly, as in the closest target image of Person 1, the deepfake may be too similar. Normalized furthest images with a randomness interval creates quantitatively dissimilar deepfakes as demonstrated in the last column. Note that in this access level, all faces except the owner are replaced, however only four of them are analyzed in detail here. Additional MFMC samples per target query type can be explored in Fig. 7.



	FTGAN	FSGAN	FaceSwap	SimSwap
PSNR	34.64	30.67	33.49	37.29
SSIM	0.90	0.89	0.93	0.97
RMSE	4.72	7.46	5.39	3.48
FID	39.72	197.64	38.66	21.55
Embed	1.09	1.18	1.16	0.99

	FTGAN	FSGAN	FaceSwap	SimSwap
PSNR	36.25	31.04	32.06	36.94
SSIM	0.96	0.91	0.93	0.97
RMSE	3.92	7.14	6.35	3.62
FID	84.62	284.65	165.79	26.45
Embed	1.07	1.20	1.16	1.01

	FTGAN	FSGAN	FaceSwap	SimSwap
PSNR	35.24	30.71	32.92	36.25
SSIM	0.90	0.83	0.86	0.94
RMSE	4.40	7.42	5.76	3.92
FID	25.13	240.93	35.60	20.98
Embed	0.95	1.04	1.11	0.92

	FTGAN	FSGAN	FaceSwap	SimSwap
PSNR	33.43	34.77	32.95	37.33
SSIM	0.85	0.96	0.92	0.96
RMSE	5.43	4.65	5.73	3.46
FID	51.33	101.14	60.32	28.61
Embed	1.32	1.27	1.24	1.00

	FTGAN	FSGAN	FaceSwap	SimSwap
PSNR	36.15	33.53	31.49	35.63
SSIM	0.95	0.92	0.93	0.96
RMSE	3.96	5.37	6.78	4.21
FID	16.96	94.15	70.26	12.26
Embed	1.11	1.29	1.27	1.19

	FTGAN	FSGAN	FaceSwap	SimSwap
PSNR	35.15	30.49	32.42	36.62
SSIM	0.94	0.87	0.89	0.95
RMSE	4.45	7.61	6.09	3.75
FID	24.27	274.40	70.35	28.15
Embed	1.15	1.14	1.17	1.03

Figure 4. **GAN Comparison.** Six sample target-source pairs identified by MFMC (left) and corresponding results by 4 GANs are compared in PSNR, SSIM, RMSE, FID, and embedding spaces.

4.1. Experiments

The quality of MFMC photos depends on the photorealism of the underlying GAN and the target image datasets. In addition, the similarity metric for target query affects the diversity and credibility of the results.

4.1.1 GAN Comparison

In Fig. 4, we compare four face generators (FTGAN [38], FSGAN [45], FaceSwap [46], and SimSwap [9]) with five quality metrics (PSNR, SSIM, and RMSE in image space; and FID and face distance in embedding space), on six MFMC results. On the left of each row, we demonstrate the target from StyleGAN dataset (top) and the source (bottom) from the party dataset, followed by the result created by each GAN and their scores. We observe that SimSwap results score as the most preferable (colored in green). Note that image and face based metrics should be evaluated together, otherwise obvious rotation/cropping artifacts as in the fourth sample may be missed. We emphasize that SimSwap creates deepfakes faithful to source resolution (row 2), source headpose and expression (row 4), target race (row 5), and target accessories (row 3). Extending this comparison to the whole dataset, we conclude that SimSwap generates the highest scoring fakes in all metrics.

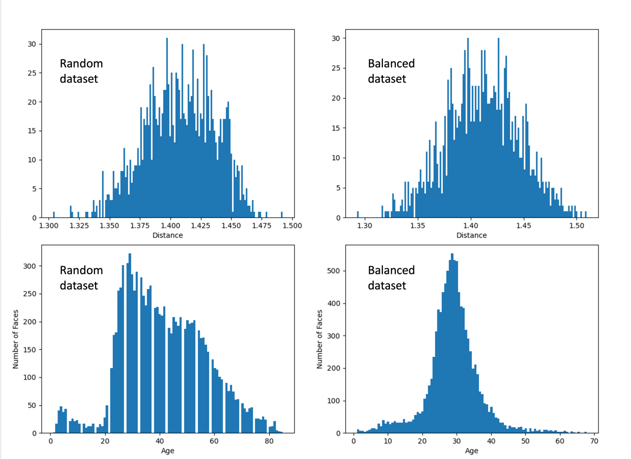


Figure 5. **Random and Balanced Datasets.** Embedding distance distribution and age distribution of the two datasets are compared.

4.1.2 Dataset Dependency

As the source repository, we utilize two different synthetic datasets. StyleGAN dataset does not have a predefined distribution of the faces, it contains randomly generated faces in terms of age, gender, and skin tones. On the other hand, Generated.photos dataset has equal distribution in terms of the ranges for the aforementioned categories. This difference is reflected in the distributions depicted in Fig. 5, and having a uniform range of attributes help MFMC create diverse samples. In contrast, distributions are more similar in terms of embedding distances, which reflects that MFMC is not affected by different datasets for target queries.

4.1.3 Similarity Metric Selection

We select a sample photo with five people to show the effects of similarity metrics for target image query. Fig. 6 gathers furthest target images per source, where the dissimilarity is defined by face embedding distance, FID, RMSE, SSIM, and PSNR in each row. To coherently compare the metrics, (1) we discard the randomness interval by using the most dissimilar and (2) we use the aforementioned equal distribution dataset. Visually and quantitatively the most dissimilar is demonstrated to be the face embedding metric in the first row, whereas FID creates a non-uniform manifold with extrema, RMSE and SSIM weighs headpose and alignment, and PSNR weighs color.

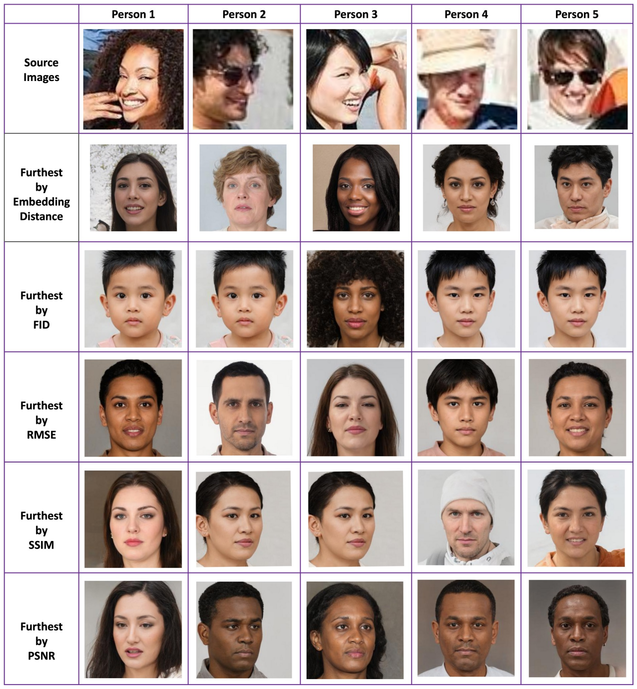


Figure 6. **Similarity Metric Comparison.** Source (first row) and queried target images per five different metrics are demonstrated.

.

4.2. Evaluation

We evaluate the success of MFMC by the reduction of overall face recognition accuracy and by visual comparison to existing privacy preserving face altercations.

4.2.1 Breaking Face Recognition

Following the main motivation of MFMC, we want to prevent mainstream face recognition systems from identifying faces. To support this motivation, we test seven state of the art face recognition systems (DeepID [59], OpenFace [6], DeepFace [60], FaceNet [53], FaceNet512 [53], DLib [30], and ArcFace [14]) on 1282 faces in 193 images created by MFMC. We use per-detector thresholds on the cosine distance between the face descriptors of source vs. target (query synthetic face) and source vs. result (created deepfake) faces from each face detector to decide if they belong to the same person. Tab. 1 documents face recognition accuracies which MFMC reduces by 61% on the average (last col.), and by 65% if we lift the randomness concern (fourth col.). The face identification accuracies between source and target faces are reported for validation (second and third cols.), these are expected to be very low as we want justifiably dissimilar deepfakes. For the most popular choice OpenFace, MFMC reduces its accuracy by 80%. We repeat this experiment using SSIM and RMSE metrics in Supp. B.

Face	Source vs. Target		Source vs. Result	
Detector	Furthest	Random	Furthest	Random
FaceNet512	0.001	0.0	0.14	0.16
OpenFace	0.001	0.002	0.2	0.23
FaceNet	0.03	0.02	0.32	0.34
DLib	0.02	0.05	0.35	0.45
ArcFace	0.06	0.04	0.36	0.45
DeepID	0.006	0.01	0.54	0.55
DeepFace	0.04	0.06	0.57	0.55
Average	0.02	0.02	0.35	0.39

Table 1. **Face Recognition Accuracies after MFMC.** Seven SOTA approaches are compared on MFMC results based on face identification accuracy, which is reduced by 61% on the average.

4.2.2 Comparison

To the best of our knowledge, MFMC is the first full system design for using deepfakes to solve face ownership problem in social media platforms. Other approaches using masking [12], filtering [17], image transformations [52], inpainting [26] and GANs [40] neither design how such algorithms can be systemized for production, nor evaluate against face recognition systems. Other approaches [50, 26, 10] do not provide control over face parameters, or extreme anonymization obfuscates the face too much that renders the photo meaningless [69, 31, 32]. To support this claim, we compare MFMC to CIAGAN [40] and Deep Privacy [26] using their sample images in Supp. A.

5. Conclusion

We present a privacy-enhancing anonymization system for everyone to have control over their faces in social photo sharing networks, saying *my face, my choice!*. In addition to the new metric for target query for deepfake creation, we extensively analyze different deepfake generators and similarity metrics for this task. MFMC also demonstrates a responsible use for deepfakes by design, especially for protecting faces of minors and vulnerable populations, in contrast to the dystopian scenarios [11]. We also present

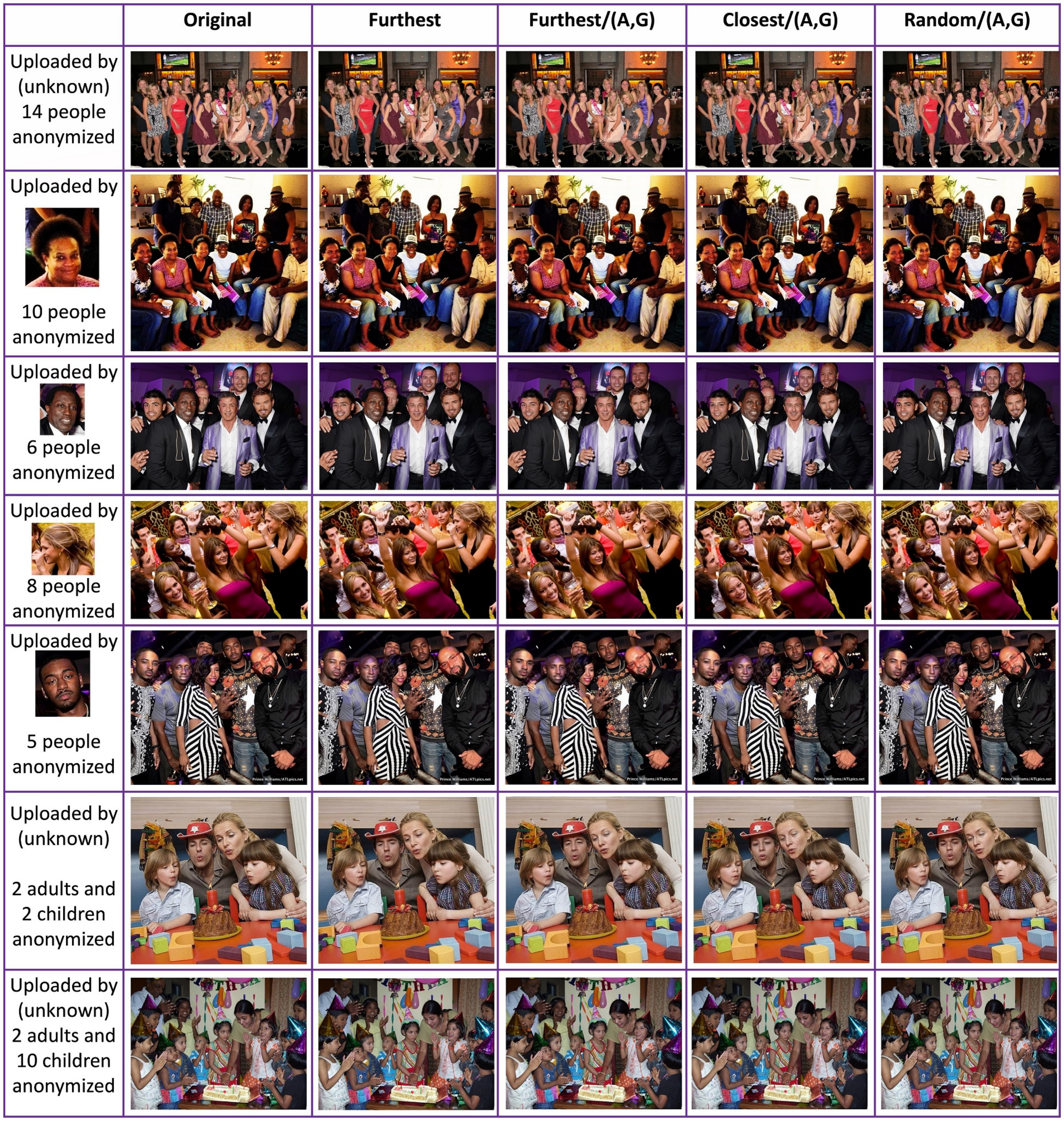

Original
Furthest
Furthest/(A,G)
Closest/(A,G)
Random/(A,G)
Uploaded by (unknown) 14 people anonymized
Uploaded by 10 people anonymized
Uploaded by 6 people anonymized
Uploaded by 8 people anonymized
Uploaded by 5 people anonymized
Uploaded by (unknown) 2 adults and 2 children anonymized
Uploaded by (unknown) 2 adults and 10 children anonymized

Figure 7. **Additional MFMC Results.** Original images and privacy enhanced versions based on different target queries are demonstrated. The last column with a randomness threshold from the furthest distance within the same age and gender shows the ultimate results.

different face access models for efficiency and embedding storage. We validate that MFMC is able confuse several current face identifications systems. We believe that current social media platforms would free the users if similar approaches to MFMC are implemented for face privacy and contextual integrity.

As a prototype system, there is always room for improvement. As mentioned in Sec. 4.1.2, the diversity and dissimilarity of the fakes depend on the distribution in the synthetic dataset for target query. The face resolution and orientation also matter, as small or oblique faces may be missed by the face detector. Finally, users with thousands of friends may consume client storage for face embeddings. It is left as future work to plan a secure client-server protocol for querying friends' face embeddings. Further discussions about our threat model and privacy evaluation are in Supp. C and D.

References

[1] Facebook, citing societal concerns, plans to shut down facial recognition system. https://www.nytimes.com/2021/11/02/technology/facebook-facial-recognition.html. Accessed: 2022-03-01.

[2] Facial recognition firm clearview ai says it will soon have 100 billion photos in its database to ensure 'almost everyone in the world will be identifiable' and wants to expand beyond law enforcement. https://www.dailymail.co.uk/news/article-10523739/Clearview-AI-seeking-100-billion-photos-facial-recognition-database.html. Accessed: 2022-03-01.

[3] A major flaw in google's algorithm allegedly tagged two black people's faces with the word 'gorillas'. https://www.businessinsider.com/google-tags-black-people-as-gorillas-2015-7. Accessed: 2022-03-01.

[4] Clearview ai. https://www.clearview.ai/, 2017.

[5] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. Past, present, and future of face recognition: A review. *Electronics*, 9(8):1188, 2020.

[6] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[7] Shai Avidan and Moshe Butman. Efficient methods for privacy preserving face detection. *Advances in neural information processing systems*, 19, 2006.

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

[9] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 2003–2011. ACM, 2020.

[10] Zhenfei Chen, Tianqing Zhu, Ping Xiong, Chenguang Wang, and Wei Ren. Privacy preservation for image data: a gan-based method. *International Journal of Intelligent Systems*, 36(4):1668–1685, 2021.

[11] David Chu, Ilke Demir, Kristen Eichensehr, Jacob G Foster, Mark L Green, Kristina Lerman, Filippo Menczer, Cailin O'Connor, Edward Parson, Lars Ruthotto, et al. White paper: Deep fakery – an action plan. Technical Report http://www.ipam.ucla.edu/wp-content/uploads/2020/01/Whitepaper-Deep-Fakery.pdf, Institute for Pure and Applied Mathematics (IPAM), University of California, Los Angeles, Los Angeles, CA, Jan. 2020.

[12] Anupam Das, Martin Degeling, Xiaoyou Wang, Junjue Wang, Norman Sadeh, and Mahadev Satyanarayanan. Assisting users in a world full of cameras: A privacy-aware infrastructure for computer vision applications. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1387–1396, 2017.

[13] Ilke Demir and Umur A Ciftci. Mixsyn: Learning composition and style for multi-source image synthesis. *arXiv preprint arXiv:2111.12705*, 2021.

[14] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[15] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.

[16] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

[17] Ádám Erdélyi, Tibor Barát, Patrick Valet, Thomas Winkler, and Bernhard Rinner. Adaptive cartooning for privacy protection in camera networks. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 44–49, 2014.

[18] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9378–9387, 2019.

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[21] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Model-based face de-identification. In *2006 Conference on computer vision and pattern recognition workshop (CVPRW'06)*, pages 161–161. IEEE, 2006.

[22] DGX Digital Identity Working Group. Digital identity in response to covid-19. https://www.tech.gov.sg/files/media/corporate-publications/FY2021/dgx_2021_digital_identity_in_response_to_covid-19.pdf, 2021. Accessed: 2022-03-01.

[23] Jia Guo and Jiankang Deng. Insightface. https://github.com/deepinsight/insightface, 2017.

[24] Aayush Gupta, Ayush Jaiswal, Yue Wu, Vivek Yadav, and Pradeep Natarajan. Adversarial mask generation for preserving visual privacy. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, 2021.

[25] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[26] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pages 565–578. Springer, 2019.

[27] Generated Media Inc. Generated photos, 2022. Online; accessed 18-February-2022.

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

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

[31] Kento Kobayashi, Keiichi Iwamura, Kitahiro Kaneda, and Isao Echizen. Surveillance camera system to achieve privacy protection and crime prevention. In *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 463–466. IEEE, 2014.

[32] Pavel Korshunov and Touradj Ebrahimi. Using face morphing to protect privacy. In *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 208–213. IEEE, 2013.

[33] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017.

[34] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri. Face recognition systems: A survey. *Sensors*, 20(2):342, 2020.

[35] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[36] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[37] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[38] Shao-An Lu. Ftgan. `https://github.com/shaoanlu/fewshot-face-translation-GAN`, 2019.

[39] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018.

[40] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2020.

[41] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 2021.

[42] Microsoft. Face api. `https://azure.microsoft.com/en-us/services/cognitive-services/face/#overview/`, 2017.

[43] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[44] Sophie Nightingale, Shruti Agarwal, Erik Härkönen, Jaakko Lehtinen, and Hany Farid. Synthetic faces: how perceptually convincing are they? *Journal of Vision*, 21(9):2015–2015, 2021.

[45] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019.

[46] Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gérard Medioni. On face segmentation, face swapping, and face perception. In *IEEE Conference on Automatic Face and Gesture Recognition*, 2018.

[47] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[48] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE, 2017.

[49] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[50] Hugo Proença. The uu-net: Reversible face de-identification for visual surveillance video footage. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):496–509, 2021.

[51] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018.

[52] Natacha Ruchaud and Jean-Luc Dugelay. Aseppi: Robust privacy protection against de-anonymization attacks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1352–1359, 2017.

[53] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[54] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1589–1604, 2020.

[55] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[56] Sola Shirai and Jacob Whitehill. Privacy-preserving annotation of face images through attribute-preserving face synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[57] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018.

[58] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 553–569, 2018.

[59] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[60] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[61] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, 73:44–51, 2016.

[62] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[63] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[64] Nishant Vishwamitra, Bart Knijnenburg, Hongxin Hu, Yifang P Kelly Caine, et al. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–47, 2017.

[65] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[66] Yunqian Wen, Li Song, Bo Liu, Ming Ding, and Rong Xie. Identitydp: Differential private identification protection for face images. *arXiv preprint arXiv:2103.01745*, 2021.

[67] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[68] Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34(1):47–60, 2019.

[69] Lin Yuan and Touradj Ebrahimi. Image privacy protection with secure jpeg transmorphing. *IET Signal Processing*, 11(9):1031–1038, 2017.

[70] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[71] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

[72] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020.

[73] Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *International Conference on Machine Learning*, pages 7614–7623. PMLR, 2019.

[74] James Zou and Londa Schiebinger. Ai can be sexist and racist—it's time to make it fair, 2018.