IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 23 November 2023, accepted 12 January 2024, date of publication 22 January 2024, date of current version 8 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3356550

RESEARCH ARTICLE

Face De-Identification Using Face Caricature

LAMYANBA LAISHRAM, JONG TAEK LEE, (Member, IEEE), AND SOON KI JUNG, (Senior Member, IEEE)

School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding author: Soon Ki Jung (skjung@knu.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government, Ministry of Science and Information and Communications Technology (MSIT), South Korea (Development of 5G-Based Predictive Visual Security Technology for Preemptive Threat Response) under Grant 2019-0-00203; and in part by MSIT under the Innovative Human Resource Development for Local Intellectualization Support Program supervised by IITP under Grant IITP-2022-RS-2022-00156389.

ABSTRACT Face privacy concerns revolve around the ethical, social, and technological implications of collecting, storing, and using facial data. With the advancement of deep learning techniques, realistic face privacy involves techniques that obscure or alter facial features effectively without compromising the usability or quality of the visual content. Modern face privacy techniques suffer from three main problems: 1) lack of human perception, 2) indistinguishability, and 3) loss of facial attributes. Modern face privacy techniques generate random, realistic faces to conceal the identifiable features of the original faces but lack the application of human perception to face de-identification. Indistinguishability arises with the highly realistic nature of fake faces used in face privacy, making it difficult to distinguish whether a face has been manipulated. Most face-privacy methods also fails to retain the facial attributes of the de-identified faces. Our face de-identification method is designed to address all three issues mentioned. We propose a novel face de-identification method that considers both human perception and face recognition models when de-identifying a face. We explore the tradeoff between a user misidentifying the original identity with a well-known celebrity and a facial recognition model that tries to identify the original identity. We generate caricature faces of the de-identified faces to ensure our manipulated faces can be distinguished effortlessly. The face caricatures are the exaggeration of the eyes and mouth region, and we provide different exaggeration scales depending on preference and application. We perform an attribute preservation optimization process to retrieve all the facial attributes. We demonstrate our method through a series of both qualitative and quantitative experiments with numerous user studies.

INDEX TERMS Face privacy, face de-identification, face caricature, human perception.

I. INTRODUCTION

Face recognition (FR) is a biometric authentication technique that uses stored data to compare and analyze face features taken from images or video frames to identify individuals. Airports and police enforcement agencies, among other institutions, have substantially used FR technology. Additionally, it is frequently used in various applications, including entertainment, access control, and security [1], [2].

Privacy concerns have increased with the growing use of FR technology. One such concern is the possibility of secretly identifying an individual based on characteristics captured from a photo or video stream. The requirement to record or save face recordings to use them for tasks involving recognition has given rise to grave concerns. Their primary cause is the possibility of security breaches or vulnerabilities in FR systems that might reveal a user's biometric data—like their face—without that user's knowledge or consent. Real-world events involving the collection of millions of facial prints over a massive network of cameras that compromised user privacy include Xinai Electronics [3]. Clearview AI is a private enterprise that collected facial photographs from social media platforms and developed a facial recognition system that matches relevant images from several databases and services [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

Concerns over privacy in biometric recognition systems have arisen as governments across the globe are taking the required steps to regulate the usage, gathering, and processing of biometric data. Australia, the United Kingdom, and the United States have made significant legal advancements in FR technology during the past ten years Smith and Miller [5]. The Japanese Act for the Protection of Personal Information came into effect in 2017 [6]. The European Union's (EU) General Data-Protection Regulation (GDPR) [7] is one of the most significant laws protecting an individual's privacy.

To avoid people's biometrics being misused, face privacy is necessary. Several studies have been done to protect the identity of faces [8], [9], [10]. Protecting a person's identity and facial features from being recognized, tracked, or used for unwanted purposes is known as "face privacy." The original face image is rendered unrecognizable through quality reduction in traditional face privacy methods. They remove any information that might reveal someone's identity by using masking [11], [12], [13], filtering [14], [15], [16], and transformation [17], [18], [19]. Using facial details for specific applications or deriving meaningful insights is difficult when using traditional methods because they deteriorate the overall quality of the images.

Instead of techniques that degrade the face-image quality, other methods can be used to preserve the privacy of face images. Face de-identification is eliminating or altering facial features that can facilitate the identification of individuals by swapping out the real person's face for a synthetic or surrogate one. The primary goal of face de-identification is to prevent the linkage of an individual's facial image to their identity, even if the image is recognized or processed. The earlier approach involves the k-same-family method [20], [21], [22], which calculates the average value of "k" within a dataset. Nevertheless, these techniques often struggle to produce detailed images while preserving essential visual qualities.

Face generation using generative AI has advanced significantly, enabling the creation of highly realistic and intricate human facial images. The generation of synthetic faces that closely mimic actual individuals has become substantially more accessible due to the utilization of advanced deep generative modeling methods like Generative Adversarial Networks (GANs) [23], [24], [25], [26]. The drawback of recent de-identification methods is that they are compelling and natural-looking, making it difficult to distinguish whether a person's face has been swapped. Creating realistic fake faces raises additional concerns, and we should be able to differentiate whether the face has been manipulated.

To further clarify the significance of distinguishability in face de-identification, let's use a surveillance system as an example. The ability to distinguish whether a face has been manipulated in surveillance footage is essential for several reasons related to security, authenticity, and the prevention of fraud. Facial recognition systems' credibility, authenticity, and dependability could be threatened if criminals could employ fake faces to avoid detection or trick the system for someone else. Unrestricted face manipulation can potentially violate people's privacy by using their facial similarities in situations they weren't a part of, which may cause many issues.

Human perception of identity refers to how individuals recognize and make sense of personal identities, distinguishing one from another through various sensory cues, cognitive processes, and contextual clues. Humans recognize others by processing facial features, unique traits, and expressions. It leverages memory and familiarity, enabling individuals to recognize people they are familiar with or have encountered. Human perception also plays a vital role in identifying an individual. Existing privacy techniques need to consider the requirements of human perception in the de-identification process.

Motivated by these limitations, we proposed a de-identification method that addresses all the above issues. Our method is designed with a focus on three goals: (i) easily distinguishable de-identified faces, (ii) the involvement of human perception in the de-identification method, and (iii) preserving attributes. We proposed a novel de-identification with face caricatures. Following the creation of face caricatures from [27], we used their pre-trained encoder and pre-trained StyleGAN for our de-identification method. We employ StyleGAN in our method as it can generate very high-resolution images and create images via style mixing. For the de-identification process, we use well-known Hollywood celebrity faces to conceal the identifiable features of the input face. The main reason for using celebrity faces is to incorporate human perception into our de-identification method and make the user think the input is some celebrity; meanwhile, a face recognition model identifies the input face. We performed several experiments, including a user study, to verify our approach. Caricature faces are generated from the de-identified faces, which makes our de-identified faces easily distinguishable. The caricatures are an exaggeration of the eyes and mouth regions. The attributes are preserved using attribute preservation losses.

This work presents several significant contributions:

1) We provide an innovative approach to face de-identification using face caricatures. The caricatured faces have exaggerated eyes and mouth areas of the face. The caricature faces make it easy for a face to distinguish whether it has been manipulated.
2) To the best of our knowledge, this is the first face de-identification method that consider human perception with face recognition models when de-identifying a face. We explore the trade-off between a user misidentifying the original identity with a well-known celebrity and a facial recognition model that tries to identify the original identity.
3) We conduct thorough qualitative and quantitative experiments, encompassing multiple user studies, to assess the quality and impact of our method.

The remainder of the paper is structured as follows: Section II provides related works on caricature, face obfuscation, and face de-identification techniques; Section III discusses our proposed de-identification method and caricature generation. Section IV details experiment settings, evaluation of results, and extensive user study; and Section V concludes this paper with current limitations and future work.

II. RELATED WORK

A. FACE CARICATURES

Crafting caricatures involves identifying and amplifying distinctive facial features while retaining the person's recognizable traits. Traditionally, methods like explicitly identifying and warping landmarks [28], [29] or using data-driven approaches to estimate unique facial attributes [30], [31] are used to increase the deviation from the average. Style transfer has been incorporated into some image-to-image translation techniques as generative networks have progressed [32], [33]. WarpGAN [34] performs shape exaggeration and visual quality, offering spatial variability flexibility for texture and image geometry. AutoToon [35] applies exaggerations via deformation fields and learns warping fields through supervised training on paired data from artist-warped photos. These paired data have significant spatial variations, which lead to lower-quality visual outputs. With the advancement of GAN, several methods started using GAN to create caricatures. CariGAN [36] is a GAN that focuses on image-to-caricature translation and is trained using unpaired images. Shi et al. [34] present a complete GAN framework that trains style and warping simultaneously. Creating realistic caricatures with exaggerated eyes and mouths in the work of [37] can be used in real-life applications. They extended the caricature creation with eyeglasses and designed a model using StyleGAN to generate caricatures of different styles and exaggerations [27].

Unlike the cartoonish nature of caricatures, our application requires a realistic nature of caricature that can easily blend into real-world scenarios and one that doesn't stand out. Our de-identification method uses caricatures inspired by [27] and [37], which is the exaggeration of eyes and mouth and is applicable in real-world applications.

B. FACE OBFUSCATION

The initial privacy-preserving methods that were put forth relied on hiding the individual's face. This indicates that various techniques, such as masking [11], [12], [13], filtering [14], [15], [16], and transformation [17], [18], [19], eliminate personally identifiable information. The face region is covered with a shape in the masking approach so that the person's face is fully hidden; filtering and transformation reduce the face region's resolution; and blurring employs Gaussian filters with different standard deviation values to allow for varying degrees of blurring. Some masking techniques use heat signatures to detect faces because they often appear as warmer items in the image [12], [38]. Wang et al. [13] showed a real-world application for enhancing privacy using face masking and blurring methods. A reversible technique for face masking was also introduced by Yuan and Ebrahimi [39].

Although these methods effectively protect identity information, they reduce image quality and practicality. Our method maintains visual fidelity, producing a natural appearance while hiding the distinctive, identifiable features of the original face.

C. FACE DE-IDENTIFICATION

Face de-identification now mostly depends on Generative Adversarial Networks (GANs) to produce better de-identified photos due to their improvement. Some methods utilized a GAN-based inpainting method for de-identification by incorporating facial landmark points to preserve the head pose while inpainting the head region, thus maintaining the overall structure and appearance of the de-identified image [40]. Reference [41] proposed a three-stage framework for image de-identification by projecting the identified private objects into a latent space and generating de-identification content using StyleGAN. De-identification in videos involves replacing the original identity with either a real identity from a different source or a synthesized identity that does not exist in reality [42], [43], [44]. The work of [43] applied deepfake technology to de-identify medical examination videos by swapping the patients' faces.

The focus of facial de-identification research has shifted recently to altering an individual's identity while maintaining characteristics that are unique to them. A few studies [44], [45], [46] minimize the cosine similarity of identity features to improve networks or latent codes for de-identification. Other methods [47], [48], [49], [50] use supplied attribute information to create new faces while hiding certain facial parts. DeepPrivacy conditionally generates anonymous photos that fulfill the face surroundings and sparse pose information [47]. CIAGAN [48] anticipates using masks, landmarks, and desired identities to regulate the created anonymous identities. Even if the works mentioned earlier can produce anonymous faces, they frequently need help with unnaturalness, lack of diversity, and poor practicality. Reversible face de-identification technology has garnered interest recently. FIT [49] trains a generative adversarial network with predefined binary passwords and face photos. The network can achieve anonymity through passwords and reconstruct the original face with inverse passwords. RiDDLE [50] maps the image to W latent space through GAN inversion. The password serves as guidance for other modalities, directing the editing of the latent code to alter the identity.

Our approach uses the cutting-edge StyleGAN [25], [51] generation capability to produce realistic faces. Our de-identification process uses a pre-trained encoder and StyleGAN from [27], which can also produce caricature faces. Unlike other de-identification methods, our de-identified

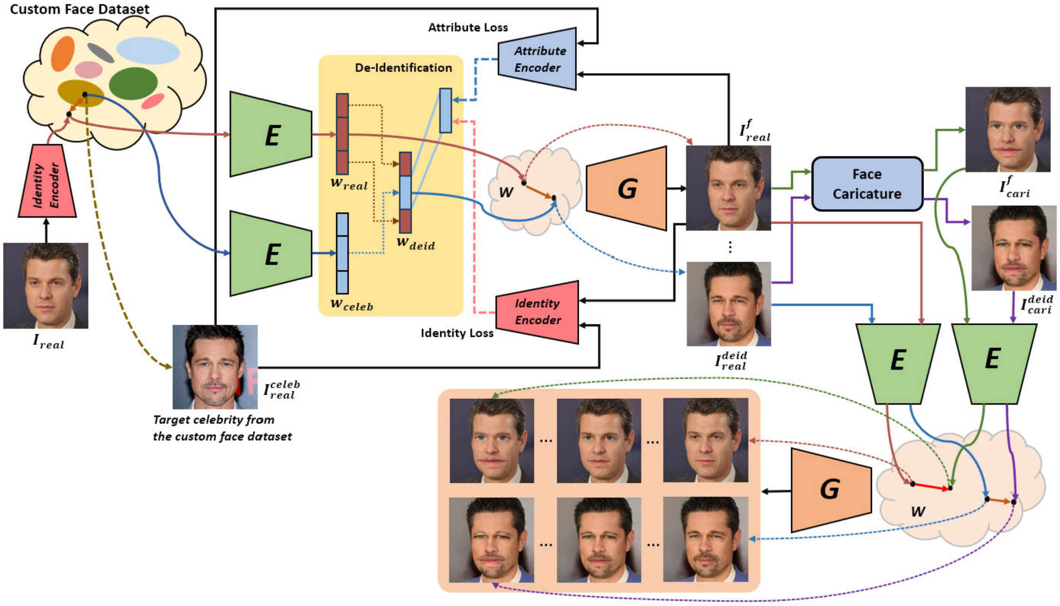

FIGURE 1. The overview of our de-identification framework. By optimizing latent code, we perform de-identification using a target celebrity face utilizing a pre-trained encoder and a StyleGAN. We generate caricatures from the incremental de-identified faces as our final result for privacy protection.

caricature faces can be easily distinguishable, whether they are manipulated or not.

III. PROPOSED DE-IDENTIFICATION

Our proposed method utilizes well-known celebrity faces as a reference for the de-identification process and is followed by creating corresponding caricature faces for the de-identified faces. We visualize our de-identification framework overview in Figure 1. Given the face image to be protected, we find the nearest celebrity face from our custom celebrity feature space. We project two latent codes using two pre-trained encoders from [27]. The latent codes are optimized to produce a de-identified face using StyleGAN latent space so that the new identity is celebrity while the pose and skin tone is from the input face. We perform an incremental process of generating a series of faces from the input face towards the de-identified faces using the StyleGAN latent space W. In each incremental step, some of the facial features of the celebrity appear to be incorporated into the input faces, in which the celebrity faces slowly replace the original facial features until the final de-identified faces are created. We create caricatures of all the incremental faces to provide additional privacy protection in our method. The caricatures are generated using two pre-trained encoders with StyleGAN latent space, similar to [27].

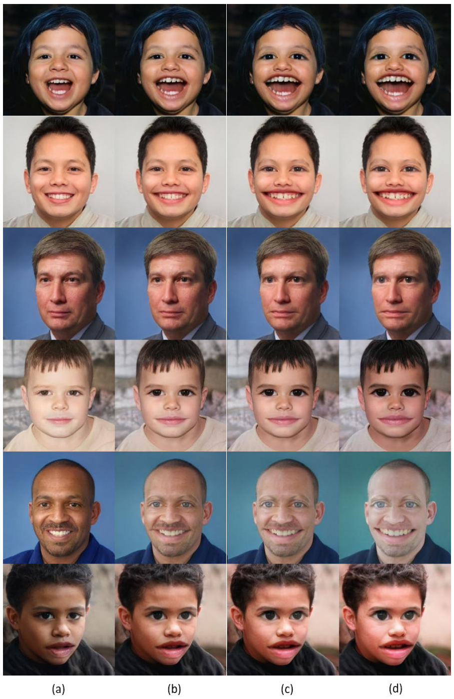

FIGURE 2. Examples of caricature faces with three different incremental exaggeration scales. (a) Input face, (b) caricature face with scale → small, (c) caricature face with scale → medium, and (d) caricature face with scale → large. Rows 1-3 represent normal caricature exaggeration, and rows 4-6 represent caricature exaggeration with different styles.

A. HIGH-QUALITY FACE CARICATURES

We utilize the creation of high-quality caricatures from real images proposed in [27], for our de-identification method. This method comprises of two stages for creating high-quality caricatures from real images: face caricature generation and face caricature projection. The face caricature generation creates new face caricature datasets, which are the exaggerated eyes and mouth regions while preserving facial contours. The face caricature datasets are created using the

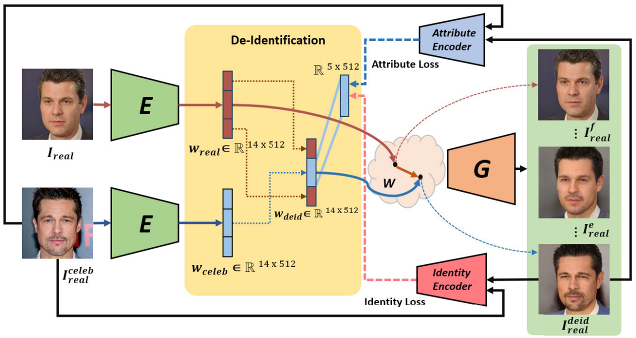

FIGURE 3. Illustration of our de-identification optimization method. Given the input and the target celebrity, two corresponding latent codes are optimized to produce a de-identified face. Several images are generated using the pre-trained StyleGAN latent space W, from the input face to the de-identified face.

FFHQ [25] and CelebA-HQ [52] datasets. They also create exaggerated faces with eyeglasses (like reading glasses and sunglasses) to enrich their caricature dataset. Using the real and new caricature datasets, they trained a StyleGAN [51] G, which can generate highly realistic images in different styles. The trained generator is capable of producing real and caricature faces with diverse facial attributes, including variations in skin tone, hair color, shapes, and more.

The face caricature projection utilizes an encoder E with the pre-trained StyleGAN G to produce high-quality caricature faces from real images. The encoder is trained using real and caricature faces while keeping the G fixed. The projecting result of the real and caricature images preserves the facial identity, attributes, and expressions from the input images. Additionally, they perform an incremental facial exaggeration from the real to the caricature images using the projected real and caricature images. Figure 2 shows the incremental exaggeration of their caricature results. They can also change style while performing incremental exaggerations. The incremental exaggeration is achieved by using two previously pre-trained encoders and projecting two latent codes on the StyleGAN latent space W, one latent from the real and the other for the corresponding caricature, and performing incremental exaggeration of caricature faces. The new face caricature datasets and the caricature projection results are 256 × 256 image resolutions.

Our de-identification method employs the pre-trained encoder E and pre-trained StyleGAN G to produce caricature faces for the de-identified faces. The projected caricature faces provide additional privacy protection for the input faces. We also utilize the incremental facial exaggeration process for caricature creation to provide an additional step on how much exaggeration is needed for a specific application. For our implementation, we divide our exaggeration into three scales: small, medium, and large, as shown in Figure 2.

B. FACE DE-IDENTIFICATION WITH CELEBRITY

Our de-identification process uses celebrity faces to protect the privacy of the input faces. First, we collected a set of well-known Hollywood celebrities. We collected 2,000 face images of 20 celebrities, making 100 faces for each celebrity. We create a 512-dimensional feature space of celebrity faces using a pre-trained image encoder [53] and Arcface [54]. We now obtain a well-defined feature representation for all the celebrity faces in our custom celebrity dataset.

Given an input image I_{real} that needs protection, we use the pre-trained network [53] to find the closest celebrity face regarding the Euclidean distance of both representations. After finding the corresponding target celebrity face I^{celeb}_{real}, we use two pre-trained encoders E from [27] to produce two corresponding latent codes, $w_{celeb} \in \mathbb{R}^{14X512}$ and $w_{real} \in \mathbb{R}^{14X512}$ where w_{celeb} is the target celebrity latent code and w_{real} is the input image latent code. Using w_{real} and w_{celeb}, we replace the middle layers of the latent code w_{real} with the middle layers of w_{celeb} to create a new latent code $w_{deid} \in \mathbb{R}^{14X512}$. The new latent code w_{deid} has the first three layers (layers 0-2) and the last layers (layers 8–13) with layers of the real latent code w_{real} and middle layers (layers 3-7) of the w_{celeb}, as shown in Figure 3. The middle five layers of w_{deid} from w_{celeb} are set as a trainable vector, and the other layers are non-trainable. We further optimized these trainable layers to get all the facial features of the target celebrity faces, discussed in Section III-B1. Our method only works on 256 × 256 resolution faces and 14 StyleGAN layers. Since each layer in our de-identification technique represents a different generation characteristic of StyleGAN output faces, layer swapping is essential. The head position, facial expression, and other coarse geometric features are preserved in layers 0–2, just as in the corresponding input faces. The target celebrity's lips, nose, eyes, and other facial characteristics are retained in layers 3–7. The color distribution and finer features, such as skin tone and hair color, are preserved in layers 8–13.

1) CELEBRITY OPTIMIZATION

To create a de-identified celebrity face, we perform an optimization process of the new latent code w_{deid}. Using the latent code w_{deid} and the pre-trained StyleGAN from [27], we can generate a new de-identified face I^{deid}_{real}. We calculate our losses using the target celebrity faces and the newly created, de-identified faces.

We use an identity loss $\mathcal{L}_{ID}(I^{deid}_{real}, I^{celeb}_{real})$ so that I^{deid}_{real} retains a similar identity to I^{celeb}_{real} and an attribute loss $\mathcal{L}_{ATT}(I^{deid}_{real}, I^{celeb}_{real})$ so that I^{deid}_{real} imposes all the facial attributes of the celebrity I^{celeb}_{real}. The celebrity optimization of w_{deid} is performed only on the middle layers (layers 3–7) using the pre-trained StyleGAN.
The identity loss is defined as follows:

$$\mathcal{L}_{ID}(I^{deid}_{real}, I^{celeb}_{real}) = | \, cos(A(I^{deid}_{real}), A(I^{celeb}_{real})) - \alpha \, |, \tag{1}$$

where $\cos(\cdot, \cdot)$ denotes the cosine distance, A denotes the ArcFace [54] network and α controls the similarity between the target celebrity and the de-identified face images.

De-identified faces with significant identity discrepancies result when $\alpha = 0$ because the identity loss enforces orthogonality between the target celebrity and the de-identified face photos. In contrast, when $\alpha = 1$, the identity loss enforces a high similarity between the de-identified face and the target celebrity. The α controls the trade-off between face usability and protecting privacy.

Attribute preservation loss is defined as follows:

$$\mathcal{L}_{ATT}(I_{real}^{deid}, I_{real}^{celeb}) = \| B(I_{real}^{deid}) - B(I_{real}^{celeb})) \|_1, \quad (2)$$

where B denotes the attribute image encoder [53]. Using flattened 512-dimensional vectors from the encoder improves attribute preservation of de-identified faces relative to the target celebrity face.

C. CARICATURE FOR DE-IDENTIFIED FACES

After performing the face de-identification optimization, we produced the final de-identified face with the target celebrity identity I_{real}^{deid}. Utilizing the latent space W of pre-trained StyleGAN G from [27], we perform the interpolation step from the w_{real} to the w_{deid} to produce multiple faces from the real input face I_{real} towards the de-identified face I_{real}^{deid}, as shown in Figure 4. Each step from I_{real} to I_{real}^{deid} creates the illusion that a minor facial feature of I_{real}^{celeb} is added to the I_{real} while at the same time removing the facial feature of the I_{real}.

We create caricatures for all the images produced in each of the steps. The caricatures are generated using pre-trained E and G, similar to those in [27]. The caricatures are produced in three scales → small, medium, and large. The scales provide the flexibility that our caricature uses for different preferences or applications. The caricatures provide additional privacy for the faces, which will be discussed in Section IV.

IV. EXPERIMENTS

A. DATASET

We perform our experiment using the Celeba-HQ [52] dataset, which contains 30,000 1024 × 1024 face images of celebrities from the CelebA dataset with various attributes like age, gender, race, etc. For our implementation, we resized all the images to 256 × 256 resolution.

B. IMPLEMENTATION

Our de-identification framework utilizes multiple pre-trained networks: a pre-trained encoder and a pre-trained Style-GAN [27], ArcFace [54], attribute image network [53]. The optimization training of the middle layers of w_{deid} is performed for 50 epochs. Our optimization is performed on a single NVIDIA TITAN Xp with 12 GB of VRAM. Each latent code optimization takes ~1 minute/epoch. The latent code generation from the encoders takes ~20 seconds. Our whole de-identification process is performed on 256 × 256 images.

C. DE-IDENTIFIED CARICATURE FACES

We compare different results with different de-identification optimization margin values for caricatures. In Figure 5, we show the qualitative results of our de-identification method with different values of the α that controls the dissimilarity between the real and the de-identified face images. When $\alpha = 1$, the de-identified face is highly similar to the target celebrity face. In contrast, when $\alpha = 0$, the de-identified face has a high identity difference with the target celebrity face, creating a new face from the orthogonality between the target celebrity and the de-identified face.

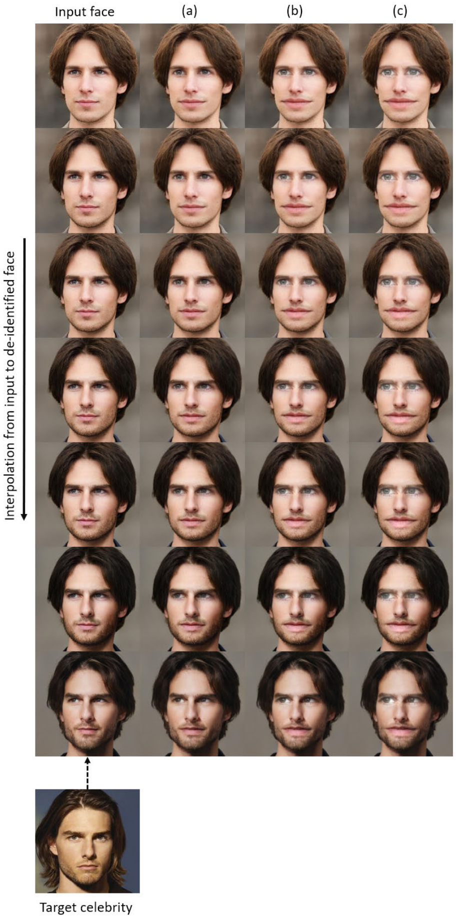

FIGURE 4. Visualization of our interpolation results from input to the de-identified face with different caricature scales. The first column presents the face from input to the de-identified face. (a) Caricature with scales → small, (b) caricature with scales → medium, and (c) caricature with scales → large.

We create different caricatures with different scales for all the de-identified faces. Figure 5 shows the caricature faces for the de-identified faces with small, medium, and large scales. The generated caricatures are distinct, and we can easily distinguish whether a face has been manipulated, which fulfills our primary goal of using face caricatures in de-identification. Different people have different preferences for facial exaggeration, so providing different scales of exaggeration (like small, medium, and large) gives flexibility in our caricature approach.

To test our different scales of caricature generation on machine recognition, we take the example shown in Figure 5.

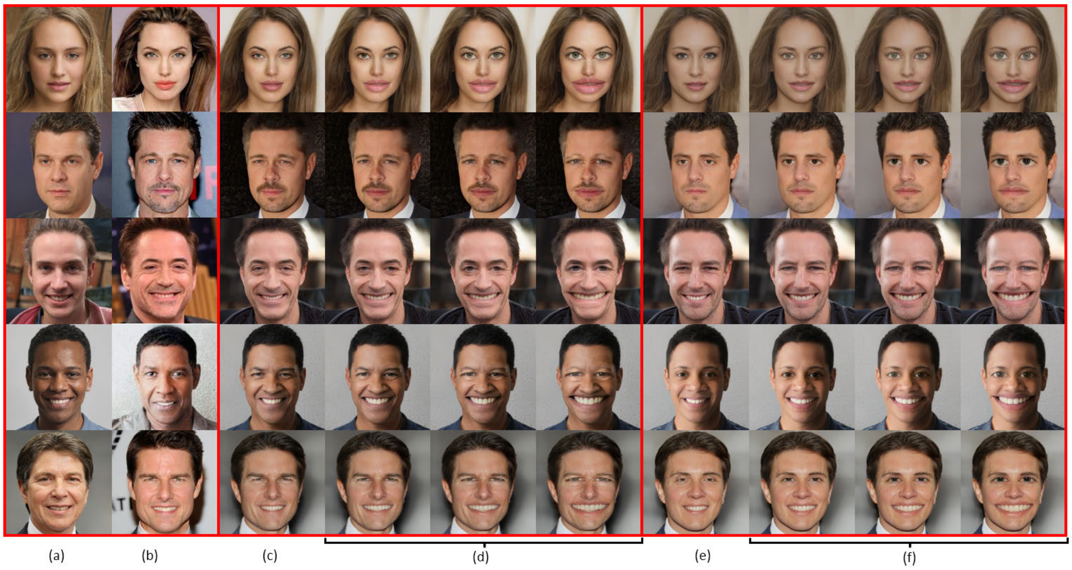

FIGURE 5. Visualization of our result with different caricature results. (a) Input face, (b) target celebrity face, (c) de-identified face [$\alpha = 1$], (d) Our caricature de-identified face [$\alpha = 1$] with different scale → [small; medium; large], (e) de-identified face [$\alpha = 0$], (f) Our caricature de-identified face [$\alpha = 0$] with different scale → [small; medium; large].

We used a face recognition (FR) model, ArcFace [54] trained with the LFW [55] and a custom face set for our face recognition experiment. Table 1 shows the recognition result. ✓ indicates when the FR successfully identifies the input face, and ✗ indicates when the FR fails to recognize the input face identity. The FR fails to recognize the caricatures with scales → large, while the FR can still recognize scale → small and medium. This shows that the scales → large provide privacy protection from trained FR.

TABLE 1. Face recognition test of the example set shown in Figure 5.

Input	(a)	(b)	(c)
✓	✓	✓	✗
✓	✓	✓	✗
✓	✓	✓	✗
✓	✓	✓	✗
✓	✓	✓	✗
✗	✗	✗	✗
✗	✗	✗	✗

D. DE-IDENTIFIED FACES

The evaluation of our de-identification results with different optimization margin values is shown in Table 2. We compare the results with the target celebrity identity. We use Curricularface [56] for computing identity similarity. The de-identified result with $\alpha = 1$ has higher identity similarity than $\alpha = 0$. We also compare the similarity of the results with the input faces using LPIPS, pixel l2 distance, and SSIM to evaluate the degree of facial information preservation.

TABLE 2. Comparison of our de-identified faces with target celebrity and input face.

Results	Comparing with the target celebrity	Comparing with the input image		
	Identity ↑	LPIPS ↓	L2 ↓	SSIM ↓
Our de-id ($\alpha = 1$)	0.58	0.24	0.18	0.56
Our de-id ($\alpha = 0$)	0.20	0.24	0.17	0.59

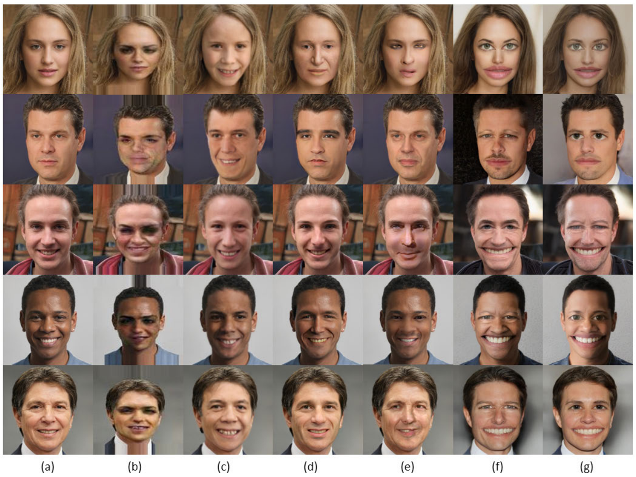

FIGURE 6. Comparison of different face privacy methods. (a) Input face, (b) CIAGAN [48], (c) DeepPrivacy [47], (d) FIT [49], (e)RiDDLE [50], (f) our caricature de-identified face [$\alpha = 1$] with scales → large, and (g) our caricature de-identified face [$\alpha = 0$] with scales → large.

E. COMPARING DIFFERENT DE-IDENTIFICATION TECHNIQUES

We compare our method with only de-identification methods, as face-obfuscation methods conceal identities but compromise the usefulness of the image. We compare our method with CIAGAN [48], DeepPrivacy [47], FIT [49] and RiDDLE [50], as shown in Figure 6. CIAGAN drastically ruins the image's original structure and has unsatisfactory artifacts. Although DeepPrivacy generates identical facial traits, it limits identity variation despite having superior image quality and realism. FIT can produce highly concealed, anonymous photos with success. In general, RiDDLE produces realistic and varied anonymous faces. The issue with these techniques is that it is very difficult to distinguish whether the face has been swapped or manipulated, as most results are very realistic. Our de-identification generates distinct caricature faces in which a person can easily distinguish between manipulated and original faces while keeping all the facial attributes.

We perform a quantitative evaluation between all the de-identification methods: CIAGAN [48], DeepPrivacy [47], FIT [49] and RiDDLE [50] with our results and different caricature results. Table 3 shows the evaluation results. We randomly sampled 500 celebA-HQ and performed all the de-identification methods for comparison. We use Curricularface [56] for computing identity similarity. It is important to note that our different caricature scales provide additional privacy protection. Scale → large provides the most protection. Table 4 shows the evaluation of the usability of our results. Our approach excels in maintaining comprehensive facial details. We assess the image quality of our result using the Fréchet Inception Distance [57]. Our technique attains a lower FID, signifying superior generation quality in deidentification compared to current methods.

TABLE 3. Comparison of identification rate with other methods. The lower the de-identification value in the table, the better the privacy protection.

Methods		Identity ↓
DeepPrivacy [47]		0.08
CIAGAN [48]		0.21
FIT [49]		0.22
RiDDLE [50]		0.16
Our de-id ($\alpha = 1$)		0.15
Our de-id caricature ($\alpha = 1$)	Scale → small	0.12
	Scale → medium	0.09
	Scale → large	0.08
Our de-id ($\alpha = 0$)		0.29
Our de-id caricature ($\alpha = 0$)	Scale → small	0.20
	Scale → medium	0.16
	Scale → large	0.15

F. USER STUDY

We perform different user studies to check how effective our method is. We perform three different user studies.

TABLE 4. Utility evaluation using dlib and MTCNN.

Method	dlib ↑	MTCNN ↑	FID ↓
DeepPrivacy [47]	96.60	99.54	28.97
Ciagan [48]	95.14	96.88	92.34
FIT [49]	99.7	99.3	28.57
RiDDLE [50]	99.1	100	29.97
Our de-id ($\alpha = 1$)	100	100	26.31
Our de-id ($\alpha = 0$)	100	100	26.82

1) USER STUDY 1: RECOGNITION RATE OF CARICATURE FACES

The first user study is to check if our caricatures are recognizable. This study aims to check whether a user can recognize a celebrity, given that the celebrity is well-known and the user is very familiar with the celebrity. We experimented with caricature exaggeration scales → small, medium, and large. We gathered 30 movie enthusiasts and provided 30 Hollywood celebrity caricature faces for each scale. All the celebrity faces in each exaggeration scale are diverse, with different poses, backgrounds, etc.

TABLE 5. User Study 1: User recognition rate with different caricature scales.

Scale → small	Scale → medium	Scale → large
98.6%	94.2%	82.1%

Table 5 shows the result of the recognition rate. We can conclude that even after scale → large exaggeration of the face, a person can easily recognize the identity of the caricature face, given that they are very familiar with the target identity.

2) USER STUDY 2: HUMAN PERCEPTION VS FACE RECOGNITION MODEL

We perform our second user study with human perception. This experiment aims to know how much human perception affects the de-identification process and compare the user recognition results with the results of a face recognition model. We use well-known Hollywood celebrity faces to trick a person into thinking that the de-identified face is the target celebrity, while the trained face recognition model recognizes the identity as the input face. An example set is shown in Figure 7, where we can see that the features of the input face appear to be more similar to the celebrity face in each step.

We used a face recognition (FR) model, ArcFace [54] trained with the LFW [55] and a custom face set for our face recognition experiment. Table 6 represents the result of the example set in Figure 7 for the face recognition model. We observe that the FR model recognizes the input faces till column (3), even though many facial features of the target celebrity face are present. ✓ indicates when the FR successfully identifies the input face, and ✗ indicates when the FR fails to recognize the input face identity. This

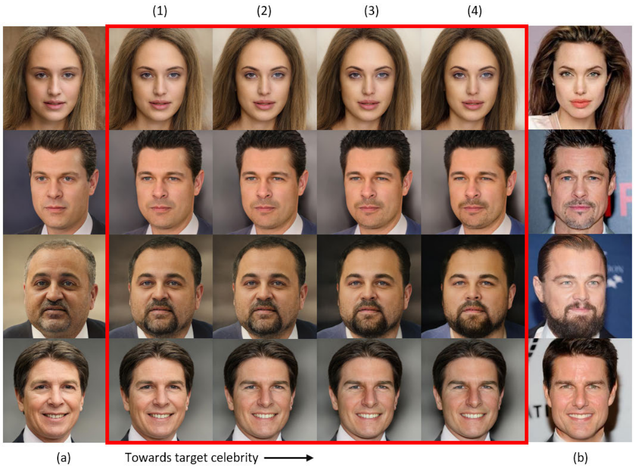

FIGURE 7. User Study 2: Examples of de-identification steps towards celebrity. (a) Input face, and (b) de-identified face.

visualizes the margin between the input and the celebrity faces in the trained FR feature embedding.

We evaluate the human perception using multiple users on the condition that the users are familiar with our selected celebrities. Table 7 represents the result of the example set in Figure 7 for the human perception experiment. We can trick a user into thinking that the identity face is a celebrity while the FR recognizes it as the input identity. The user still thinks the identity of the face is a celebrity till some cases in column (2). ✓ indicates when the user recognizes the face as a celebrity, and ✗ indicates when the user doesn't recognize the face as a celebrity.

The selection of which face to use for the de-identification depends on the specific application. There is a tradeoff when selecting the de-identified face depending on the application, whether we want the misclassification for an FR system or to trick user perception. Our approach provides a wide range of faces for de-identifying one specific face.

TABLE 6. User Study 2: Face recognition result of input face from Figure 7.

(1)	(2)	(3)	(4)
✓	✓	✗	✗
✓	✓	✓	✗
✓	✓	✓	✗
✓	✓	✓	✓

TABLE 7. User Study 2: User recognizing as celebrity from Figure 7.

(1)	(2)	(3)	(4)
✗	✗	✓	✓
✗	✓	✓	✓
✗	✗	✓	✓
✗	✓	✓	✓

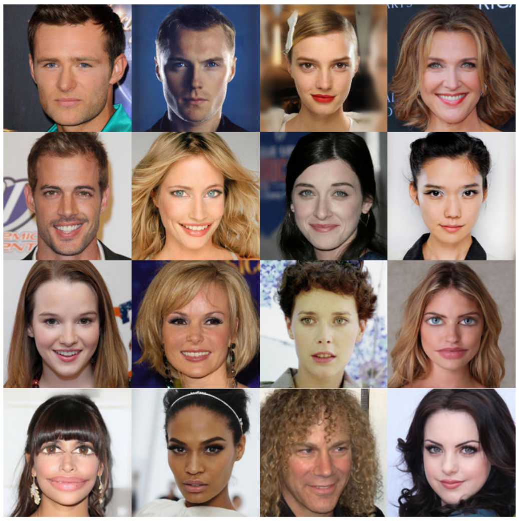

FIGURE 8. User Study 3: Examples set of the user study on distinguishability of caricature faces. Each row represent one example set.

TABLE 8. User Study 3: User distinguishable rate for manipulating faces using different caricature scales.

Scale → small	Scale → medium	Scale → large
61.7%	94.8%	100%

3) USER STUDY 3: DISTINGUISHABILITY OF CARICATURE FACES

We performed our third user study on caricature distinguishability. This experiment aims to check how easily a caricature face can be distinguished, provided the caricature faces are manipulated. We experimented with caricature exaggeration scales → small, medium, and large. We gathered 30 users for our experiment. Figure 8 shows an example of our experimental setup. Each row provides one specific example and consists of four faces. We also provide face sets with no caricature, as shown in the first row, a set of four original images with no caricature. The second row is an example for scale → small, the third row is for scale → medium, and the fourth row is for scale → large. We experimented with a total set of 80, each with four face images. The sets are uniformly distributed into 20 sets each, 20 sets of no caricature, 20 sets of caricature with scale → small, 20 sets of caricature set with a scale → medium, and 20 sets of caricature with a scale → large. We mixed the sets for our experiment.

Table 8 represents the user distinguishable rate for our caricature with three different scales. We observe that with scale → large, it is effortless to distinguish caricature faces from original faces. Most of the caricatures with scale → medium can be identified. To distinguish caricature faces with scale → small scale is difficult, as there are faces with big eyes and lips by nature, creating confusion during the experiment.

V. CONCLUSION

We provide a novel method for face de-identification using facial caricatures. The caricatures are the exaggeration of the mouth and the eye regions of the faces, which makes them distinct. The caricature makes the faces easily distinguishable, whether they have been altered or not. As far as we are aware, this is the first face de-identification technique that takes into account both face recognition models and human perception. We investigate the trade-off between a facial recognition model that attempts to identify the original identity and a user who incorrectly associates the original identity with a well-known celebrity. We evaluate the effectiveness and impact of our method by performing extensive user research and qualitative and quantitative evaluations.

A. LIMITATION AND FUTURE WORK

Our de-identification framework depends on a pre-trained StyleGAN generator, which introduces particular constraints inherent in the framework due to the limitations of the utilized GAN generator in reproducing faces that statistically resemble the originals. Another area for improvement within the proposed framework is the inversion method, which could potentially result in inaccurate latent code inversions that could consequently impact the de-identification outcomes. For the future work, the inversion process can be improved with the aspect of facial occlusion and also the potential to process faster and more accurately in multiple video frames. Improvement of the de-identification process for the target celebrity using additional losses and changes in the framework to make it more robust.

REFERENCES

[1] *Facial Recognition: Do You Really Control How Your Face is Being Used? USA Today*, 2019. [Online]. Available: https://www.usatoday.com/story/tech/2019/11/19/police-technology-and-surveillance-politics-of-facial-recognition/4203720002/

[2] I. Global. (2021). *Role of CCTV Cameras: Public, Privacy and Protection*. [Online]. Available: https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

[3] Z. Whittaker. (2022). *A Huge Chinese Database of Faces and Vehicle License Plates Spilled Online*. [Online]. Available: https://techcrunch.com/2022/08/30/china-database-face-recognition/

[4] (2020). *The Secretive Company That Might End Privacy as We Know it. New York Times*. [Online]. Available: https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

[5] M. Smith and S. Miller, "The ethical application of biometric facial recognition technology," *AI Soc.*, vol. 37, no. 1, pp. 167–175, Mar. 2022.

[6] NEC. (2018). *Privacy Measures of Biometrics Businesses*. [Online]. Available: https://www.nec.com/en/global/techrep/journal/g18/n02/180205.html

[7] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (GDPR)," in *A Practical Guide*, vol. 10, 1st, Ed. Cham, Switzerland: Springer, 2017, p. 5555.

[8] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022.

[9] A. Chadha, V. Kumar, S. Kashyap, and M. Gupta, "Deepfake: An overview," in *Proc. 2nd Int. Conf. Comput., Commun., Cyber-Secur.* Cham, Switzerland: Springer, 2021, pp. 557–566.

[10] L. Laishram, M. M. Rahman, and S. K. Jung, "Challenges and applications of face deepfake," in *Proc. Int. Workshop Frontiers Comput. Vis.* Cham, Switzerland: Springer, 2021, pp. 131–156.

[11] N. Babaguchi, T. Koshimizu, I. Umata, and T. Toriyama, "Psychological study for designing privacy protected video surveillance system: PriSurv," in *Protecting Privacy in Video Surveillance*. Berlin, Germany: Springer, 2009, pp. 147–164.

[12] Y. Zhang, Y. Lu, H. Nagahara, and R.-I. Taniguchi, "Anonymous camera for privacy protection," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4170–4175.

[13] J. Wang, B. Amos, A. Das, P. Pillai, N. Sadeh, and M. Satyanarayanan, "A scalable and privacy-aware IoT service for live video analytics," in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 38–49.

[14] H. Fradi, V. Eiselein, I. Keller, J.-L. Dugelay, and T. Sikora, "Crowd context-dependent privacy protection filters," in *Proc. 18th Int. Conf. Digit. Signal Process. (DSP)*, 2013, pp. 1–6.

[15] P. Korshunov, C. Araimo, F. De Simone, C. Velardo, J.-L. Dugelay, and T. Ebrahimi, "Subjective study of privacy filters in video surveillance," in *Proc. IEEE 14th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2012, pp. 378–382.

[16] L. Fan, "Differential privacy for image publication," in *Proc. Theory Pract. Differ. Privacy (TPDP) Workshop*, 2019, vol. 1, no. 2, p. 6.

[17] S. Dadkhah, M. Koeppen, S. Sadeghi, and K. Yoshida, "Bad AI: Investigating the effect of half-toning techniques on unwanted face detection systems," in *Proc. 9th IFIP Int. Conf. New Technol., Mobility Secur. (NTMS)*, Feb. 2018, pp. 1–5.

[18] K. Kobayashi, K. Iwamura, K. Kaneda, and I. Echizen, "Surveillance camera system to achieve privacy protection and crime prevention," in *Proc. 10th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Aug. 2014, pp. 463–466.

[19] L. Fan, "Practical image obfuscation with provable privacy," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 784–789.

[20] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 232–243, Feb. 2005.

[21] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[22] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *Proc. Int. Workshop Privacy Enhancing Technol.* Cham, Switzerland: Springer, 2005, pp. 227–242.

[23] M. A. Diniz and W. R. Schwartz, "Face attributes as cues for deep face recognition understanding," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 307–313.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[25] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[26] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[27] L. Laishram, M. Shaheryar, J. T. Lee, and S. K. Jung, "High-quality face caricature via style translation," *IEEE Access*, vol. 11, pp. 138882–138896, 2023.

[28] B. Gooch, E. Reinhard, and A. Gooch, "Human facial illustrations: Creation and psychophysical evaluation," *ACM Trans. Graph.*, vol. 23, no. 1, pp. 27–44, Jan. 2004.

[29] P.-Y. C. W.-H. Liao and T.-Y. Li, "Automatic caricature generation by analyzing facial features," in *Proc. Asia Conf. Comput. Vis. (ACCV)*, vol. 2, 2004, p. 2.

[30] J. Liu, Y. Chen, and W. Gao, "Mapping learning in eigenspace for harmonious caricature generation," in *Proc. 14th ACM Int. Conf. Multimedia*, 2006, pp. 683–686.

[31] Y. Zhang, W. Dong, C. Ma, X. Mei, K. Li, F. Huang, B.-G. Hu, and O. Deussen, "Data-driven synthesis of cartoon faces using different styles," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 464–478, Jan. 2017.

[32] Z. Zheng, C. Wang, Z. Yu, N. Wang, H. Zheng, and B. Zheng, "Unpaired photo-to-caricature translation on faces in the wild," *Neurocomputing*, vol. 355, pp. 71–81, Aug. 2019.

[33] W. Li, W. Xiong, H. Liao, J. Huo, Y. Gao, and J. Luo, "CariGAN: Caricature generation through weakly paired adversarial learning," *Neural Netw.*, vol. 132, pp. 66–74, Dec. 2020.

[34] Y. Shi, D. Deb, and A. K. Jain, "WarpGAN: Automatic caricature generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 10762–10771.
[35] J. Gong, Y. Hold-Geoffroy, and J. Lu, "Autotoon: Automatic geometric warping for face cartoon generation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 360–369.
[36] K. Cao, J. Liao, and L. Yuan, "CariGANs: Unpaired photo-to-caricature translation," 2018, *arXiv:1811.00222*.
[37] L. Laishram, M. Shaheryar, J. T. Lee, and S. K. Jung, "A style-based caricature generator," in *Proc. Int. Workshop Frontiers Comput. Vis.* Cham, Switzerland: Springer, 2023, pp. 71–82.
[38] J. Schiff, M. Meingast, D. K. Mulligan, S. Sastry, and K. Goldberg, "Respectful cameras: Detecting visual markers in real-time to address privacy concerns," in *Protecting Privacy in Video Surveillance*. Berlin, Germany: Springer, 2009, pp. 65–89.
[39] L. Yuan and T. Ebrahimi, "Image privacy protection with secure JPEG transmorphing," *IET Signal Process.*, vol. 11, no. 9, pp. 1031–1038, Dec. 2017.
[40] Q. Sun, L. Ma, S. J. Oh, L. Van Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5050–5059.
[41] J. Yu, H. Xue, B. Liu, Y. Wang, S. Zhu, and M. Ding, "GAN-based differential private image privacy protection framework for the Internet of Multimedia Things," *Sensors*, vol. 21, no. 1, p. 58, Dec. 2020.
[42] B. Samarzija and S. Ribaric, "An approach to the de-identification of faces in different poses," in *Proc. 37th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, 2014, pp. 1246–1251.
[43] B. Zhu, H. Fang, Y. Sui, and L. Li, "Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 414–420.
[44] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9377–9386.
[45] S. Barattin, C. Tzelepis, I. Patras, and N. Sebe, "Attribute-preserving face dataset anonymization via latent code optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8001–8010.
[46] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 620–636.
[47] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *Proc. Int. Symp. Vis. Comput.*, Lake Tahoe, NV, USA. Cham, Switzerland: Springer, 2019, pp. 565–578.
[48] M. Maximov, I. Elezi, and L. Leal-Taixé, "CIAGAN: Conditional identity anonymization generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5446–5455.
[49] X. Gu, W. Luo, M. S. Ryoo, and Y. J. Lee, "Password-conditioned anonymization and deanonymization with face identity transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 727–743.
[50] D. Li, W. Wang, K. Zhao, J. Dong, and T. Tan, "RiDDLE: Reversible and diversified de-identification with latent encryptor," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 8093–8102.
[51] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12104–12114.
[52] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[53] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 18697–18709.
[54] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.
[55] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, 2008, pp. 7–49.
[56] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: Adaptive curriculum learning loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5901–5910.
[57] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6629–6640.

LAMYANBA LAISHRAM received the bachelor's degree in computer science and engineering from Visvesvaraya Technological University (VTU), Bengaluru, India, in 2014, and the M.Tech. degree in computer science and engineering from Christ University, Bengaluru, in 2017. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Kyungpook National University, Daegu, South Korea. His research interests include face de-identification, face swapping, face privacy, and surveillance systems.

JONG TAEK LEE (Member, IEEE) received the B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2005, and the M.S. and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2007 and 2012, respectively. From 2012 to 2022, he was a Senior Researcher with the Electronics and Telecommunications Research Institute, South Korea. Since 2022, he has been an Assistant Professor with the School of Computer Science and Engineering, Kyungpook National University, Daegu, South Korea. His research interests include computer vision, video surveillance, and artificial intelligence in medicine.

SOON KI JUNG (Senior Member, IEEE) received the Ph.D. degree in computer science from KAIST, in 1997. From 1997 to 1998, he was a Research Associate with the University of Maryland Institute for Advanced Computer Studies (UMIACS). Since 1998, he has been with the School of Computer Science and Engineering, Kyungpook National University (KNU), Daegu, South Korea, where he is currently a Professor. From 2001 to 2002, he was a Research Associate, and from 2008 to 2009, he was a Visiting Faculty with the IRIS Computer Vision Laboratory, University of Southern California. He is the author of over 200 articles on computer vision and graphics. He holds more than 20 patents deriving from his research. His research interests include improving the understanding and performance of intelligent vision systems and VR/AR systems, mainly through the application of 3D computer vision, computer graphics, visualization, and HCI. He serves as the Vice President for the Korean Computer Graphics Society, the Korean HCI Society, and the Korean Multimedia Society.