

Forbes: Face Obfuscation Rendering via Backpropagation Refinement Scheme

Jintae Kim¹, Seungwon Yang², Seong-Gyun Jeong², and Chang-Su Kim¹

¹ School of Electrical Engineering, Korea University, Seoul, Korea

² 42dot Inc.

jtkim@mcl.korea.ac.kr, seungwon.yang@42dot.ai, seonggyun.jeong@42dot.ai,
changsukim@korea.ac.kr

Abstract. A novel algorithm for face obfuscation, called Forbes, which aims to obfuscate facial appearance recognizable by humans but preserve the identity and attributes decipherable by machines, is proposed in this paper. Forbes first applies multiple obfuscating transformations with random parameters to an image to remove the identity information distinguishable by humans. Then, it optimizes the parameters to make the transformed image decipherable by machines based on the back-propagation refinement scheme. Finally, it renders an obfuscated image by applying the transformations with the optimized parameters. Experimental results on various datasets demonstrate that Forbes achieves both human indecipherability and machine decipherability excellently. The source codes are available at <https://github.com/mcljtkim/Forbes>.

Keywords: Face obfuscation · Privacy-enhancing technologies

1 Introduction

Deep learning networks are widely used for various vision tasks [7, 15, 37]. To train such networks effectively, a huge amount of high-quality data is required in general. Hence, various datasets have been proposed, but the privacy of individuals included in the datasets should be protected [40]. Fig. 1 shows some images in popular datasets [11, 17, 31, 41], containing faces. Although privacy issues can be raised for any images, they occur mostly for facial images. As illustrated in the bottom row of Fig. 1, we obfuscate facial regions for privacy protection.

Facial image datasets are used for diverse tasks, such as face verification [6, 16, 30], gender classification [13], race classification [35], age estimation [18, 20, 36], and facial expression recognition [19, 21, 45]. Recently, to free such datasets from privacy concerns, privacy-enhancing technologies have been developed. For example, there are face anonymization methods [1, 5, 12, 33, 38] based on generative models, which replace original faces with fake ones to make the individuals' identities unrecognizable by both humans and machines. In contrast to face anonymization, face obfuscation [23–25, 43] attempts to retain the identities of individuals for machines, as illustrated in Fig. 2. An obfuscated image should be decipherable by deep learning networks but unrecognizable by humans. In



Fig. 1: Top, facial images are selected from the Vimeo-90K [41], DAVIS 2016 [31], LFW [11], CelebA [27] datasets. Bottom, the facial regions are obfuscated by the proposed Forbes algorithm.

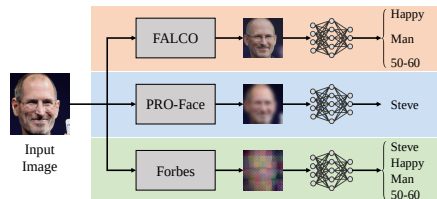


Fig. 2: Illustration of differences between the existing attribute-preserving face anonymization algorithm [1], the existing face obfuscation algorithm [43], and the proposed Forbes algorithm.

other words, an obfuscated image should satisfy two properties: human indecipherability (HI) and machine decipherability (MD).

In this paper, we propose a novel face obfuscation algorithm, called **F**ace obfuscation rendering via **b**ackpropagation **r**efinement **s**cheme (Forbes), which achieves both HI and MD. As in Fig. 2, whereas the existing obfuscation methods [23–25, 43] retain only identity information for machines, Forbes preserves attributes as well. We first develop eight parameterized obfuscating transformations. Then, we apply those transformations with the parameters to an input image to remove the identity information recognizable by humans. Then, we refine the parameters to make the transformed image decipherable by machines. Specifically, to balance HI and MD, we define energy functions for face obfuscation, minimize them using the backpropagation refinement scheme (BRS) [14], and obtain optimal parameters. We then generate an obfuscated image by applying the transformations with the optimal parameters. Extensive experiments demonstrate that the proposed Forbes achieves both HI and MD excellently.

This work has the following major contributions:

- To the best of our knowledge, Forbes is the first attempt to encrypt both identity and facial attribute information, decipherable by machines but unrecognizable by humans.
- We introduce two HI energy functions and two MD energy functions and minimize them based on BRS to strike a tradeoff between HI and MD.
- Forbes provides excellent obfuscating results on various benchmark datasets for face verification, age estimation, gender classification, and facial attribute classification.

2 Related Work

2.1 Face Anonymization

Face anonymization is a vision task to remove the identity of a person from the image. It can be achieved by simple filters, such as blurring, masking, or mosaicking. Early anonymization methods [2, 9] are based on these filters. Specifically,

Hassan *et al.* [9] masked privacy-sensitive regions, including faces, with clip art. Also, Brkić *et al.* [2] blurred a person’s face for identity removal but maintained the person’s pose and clothes. Filtering methods, however, have a limitation in that their anonymized images may be restored by inverse filters [22, 42].

Recent anonymization methods [1, 3–5, 12, 28, 32, 33, 38] have adopted generative models to synthesize identity-removed faces. For instance, Ren *et al.* [32] proposed a video anonymizer, which removes the identities of people while preserving their actions. Also, Chen *et al.* [3] incorporated a generative model into an auto-encoder for anonymization, which preserves facial expressions (*e.g.* happiness and surprise). Cho *et al.* [4] developed an auto-encoder to decompose feature vectors into identity features and attribute features (*e.g.* gender and race). They modified the identity features but left the attribute features unaltered. Thus, their method can perform attribute-preserving face anonymization. Barattin *et al.* [1] proposed another attribute-preserving anonymization method via latent code optimization. Since facial attributes are useful in various applications, such as expression recognition [45] and age estimation [20, 36], we also attempt to preserve facial attributes in the proposed obfuscation algorithm.

2.2 Face Obfuscation

Face obfuscation encrypts the identity of a person in an image so that only a machine can decrypt the identity (*i.e.* MD) while humans cannot recognize it (*i.e.* HI). Note that obfuscation is different from anonymization in that the former aims to retain some identity information decipherable by machines only, whereas the latter tries to remove all identity information and make it indecipherable by both humans and machines, as in Fig. 2. Thus, face obfuscation is essential in some applications, such as re-identification with privacy protection. However, relatively few algorithms have been developed for face obfuscation.

In an early study, Vishwamitra *et al.* [39] focused on HI only and suggested that masking a person with a black box is more effective than blurring. Li *et al.* [24] first adopted an auto-encoder for face obfuscation. They separated identity features from attribute ones, as in [4], and then performed obfuscation by replacing the attributes with different person’s attributes. Li *et al.* [23] also proposed adaptive attribute obfuscation method. They extracted facial attributes, identified important attributes for MD, and modified less important ones. However, since they focused on MD only, their obfuscated images are often recognizable by humans as well. To improve HI, Yuan *et al.* [43] proposed a deep fusion network. They first pre-obfuscated an image using pre-defined operations, such as blurring or pixelating, making it difficult for humans to recognize the identity. Then, they fused the pre-obfuscated image with the original image to yield a final obfuscation result. Although their algorithm achieves a high accuracy when tested by a face verification network, their obfuscated image still retains facial appearance decipherable by humans, as shown in Fig. 2. In contrast, the proposed Forbes algorithm makes the resultant image much less decipherable by humans. Furthermore, Forbes is attribute-preserving; from its obfuscated image, a machine can decipher facial attributes as well as identity.

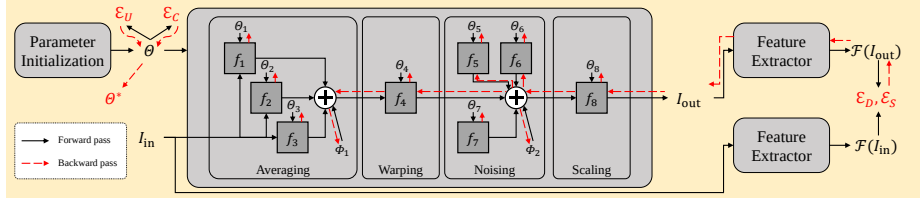


Fig. 3: An overview of the proposed Forbes algorithm. Each obfuscating transformation f_n has a set Θ_n of parameters. Also, Φ_1 and Φ_2 are composed of superposing parameters. A randomly initialized set $\Theta = \Theta_1 \cup \dots \cup \Theta_8 \cup \Phi_1 \cup \Phi_2$ is optimized to Θ^* in the backward pass to balance HI and MD. To this end, energy functions $\mathcal{E}_U$, $\mathcal{E}_C$, $\mathcal{E}_D$, and $\mathcal{E}_S$ are minimized.

3 Proposed Algorithm

Forbes uses several transformations to obfuscate an image. Each transformation has parameters to control the balance between HI and MD. We apply these transformations to process an image, after initializing their parameters randomly within their ranges. The transformations are designed, in general, to make an output image not easily recognizable by humans. However, the output may be also indecipherable by machines, such as face verification networks [16, 30]. To improve the MD, we optimize the transformation parameters to minimize the distance and maximize the cosine similarity between the two feature vectors extracted from the input and output images, respectively, by a feature extractor. For the optimization, we adopt the BRS technique [14]. Note that the parameters of the feature extractor are fixed during the optimization process. Finally, we render an obfuscated image using the transformations with the optimal parameters.

Fig. 3 is an overview of the Forbes algorithm. Let us describe each algorithm component subsequently.

3.1 Obfuscating Transformations

We design eight face obfuscating transformations f_n , which are listed in Table 1, in four categories: 1) averaging, 2) warping, 3) noising, and 4) scaling. Each transformation f_n has parameters. Let Θ_n denote the set of parameters for f_n . Fig. 4 shows examples of transformed images.

Averaging: In the averaging category, we process an input RGB image $I_{in} \in [0, 1]^{H \times W \times 3}$ to generate a degraded image $I_{deg} \in [0, 1]^{H \times W \times 3}$ that retains low-frequency information of the input image while removing its high-frequency information. For example, in *Mosaicking* (f_1), we divide I_{in} into $M \times N$ blocks of resolution $\frac{H}{M} \times \frac{W}{N}$ and fill in each block with its average value to yield I_{deg} . Similarly, *Horizontal Mean* (f_2) and *Vertical Mean* (f_3) fill in each pixel with the average value of the W horizontal pixels and the H vertical pixels, respectively. Let B_{in} and B_{deg} denote a block in I_{in} and its corresponding block in

Table 1: Specification of the eight obfuscating transformations: θ -range denotes the range of each parameter θ in Θ_n , and $|\Theta_n|$ is the number of parameters in Θ_n . Also, λ is a margin for the HI energy.

Category	Transformation	θ -range	$ \Theta_n $	λ
Averaging	f_1 Mosaicking	$[1, 1]$	$M \times N \times 3$	0
	f_2 Horizontal Mean	$[1, 1]$	$M \times N \times 3$	0
	f_3 Vertical Mean	$[1, 1]$	$M \times N \times 3$	0
Warping	f_4 Warping	$[-0.3, 0.3]$	$(M - 1) \times (N - 1) \times 2$	0.05
Noising	f_5 Sinusoid	$[0, \pi]$	$M \times N \times 3$	0
	f_6 Checkerboard	$[1, 1]$	$M \times N \times 3$	0
	f_7 Speckle	$[-0.5, 0.5]$	$M \times N \times 3$	0.1
Scaling	f_8 Color Scaling	$[10/11, 11/10]$	$M \times N \times 3$	1.05

I_{deg} , respectively. Then, we obtain an output block by

$$B_{\text{out}}^c = (1 - \theta^c) \cdot B_{\text{in}}^c + \theta^c \cdot B_{\text{deg}}^c \quad (1)$$

where c denotes one of the three RGB channels, and θ^c is a blending parameter. To remove the high-frequency information effectively, we fix the blending parameters to 1, as specified in Table 1.

Warping: To modify I_{in} geometrically, we adopt a geometric transformation, *Warping* (f_4). We use the $(M - 1) \times (N - 1)$ inner corners of the $M \times N$ blocks as grid points and perform grid warping to distort I_{in} . Each grid point is moved by $(\theta_x \Delta_x, \theta_y \Delta_y)$, where Δ_x and Δ_y are the horizontal and vertical sizes of the block. Both warping parameters θ_x and θ_y are selected from $[-0.3, 0.3]$ to prevent grid folding.

Noising: In the noising category, we obtain an output image by

$$I_{\text{out}} = I_{\text{in}} + \mathcal{N} \quad (2)$$

where $\mathcal{N}$ is a noise map. Similar to f_1 , we first divide I_{in} into $M \times N$ blocks. Then, in *Sinusoid* (f_5), for each block, we generate a sinusoidal noise along an axis at a random angle θ^c . The period of the sinusoid is a quarter of $\min\{\Delta_x, \Delta_y\}$, and the amplitude is 0.2. In *Checkerboard* (f_6), for each block, a 4×4 checkerboard pattern is used as noise, as shown in Figure 4. The absolute magnitude of the pattern is $0.3 \times \theta^c$ for each color channel, and the mean value is 0. In *Speckle* (f_7), we assign parameter θ^c to the center pixel of each block, and then bilinearly interpolate the remaining pixels from the center ones. Thus, the high-frequency components in an output image I_{out} may be strengthened, for the noise map $\mathcal{N}$ is composed of fluctuating patterns.

Scaling: For obfuscation, we also modulate colors with *Color Scaling* (f_8). Specifically, we obtain an output block by

$$B_{\text{out}}^c = \theta^c \cdot B_{\text{in}}^c \quad (3)$$

where θ^c is a scaling parameter. Since f_8 just modifies the scales of the color components, high-frequency information in an input image I_{in} can be preserved in the output image I_{out} .



Fig. 4: Examples of transformed images by f_n .

3.2 Composite Transformation

We combine the eight obfuscating transformations f_n , $1 \leq n \leq 8$, into a single composite transformation. Note that, when only one obfuscating transformation is used as in Fig. 4, the output image may contain some identity information still recognizable by humans, or erase most of the information decipherable by machines. To overcome this issue, we superpose the output images of the transformations in the same category.

First, we apply f_1 , f_2 , and f_3 to the input image and superpose the three output images in a blockwise manner. Specifically, let B_1 , B_2 , and B_3 denote a triplet of corresponding blocks in these output images. We normalize random parameters $\phi_1^c, \phi_2^c, \phi_3^c \in \Phi_1$ into $\bar{\phi}_1^c, \bar{\phi}_2^c$, and $\bar{\phi}_3^c$ using the softmax operator and obtain the convex combination

$$B_{\text{avg}}^c = \bar{\phi}_1^c \cdot B_1^c + \bar{\phi}_2^c \cdot B_2^c + \bar{\phi}_3^c \cdot B_3^c \quad (4)$$

where c is the RGB channel index. I_{avg} comprises these superposed blocks. Then, we apply f_4 to I_{avg} to yield a warped image I_{warped} . Since I_{warped} contains only low-frequency information, we generate high-frequency information with the noising transformations to enable a machine to decipher the identity. Similarly to (4), we normalize random parameters $\phi_5^c, \phi_6^c, \phi_7^c \in \Phi_2$ into $\bar{\phi}_5^c, \bar{\phi}_6^c$, and $\bar{\phi}_7^c$, and superpose the noise blocks with the warped block by

$$B_{\text{noi}}^c = B_{\text{warped}}^c + \bar{\phi}_5^c \cdot B_5^c + \bar{\phi}_6^c \cdot B_6^c + \bar{\phi}_7^c \cdot B_7^c \quad (5)$$

where B_n^c is a noise block from f_n . I_{noi} is composed of these superposed blocks. Last, we apply the scaling transformation f_8 to I_{noi} to yield the final output image I_{out} .

3.3 Parameter Initialization

Let Θ denote the set of all parameters for the composite transformation. It consists of four kinds of parameters;

$$\Theta = \Theta_F \cup \Theta_U \cup \Theta_C \cup \Phi \quad (6)$$

where Θ_F is the set of fixed parameters, Θ_U is the set of uniform parameters, Θ_C is the set of color parameters, and Φ is the set of composing parameters. Note that $\Theta_F = \Theta_1 \cup \Theta_2 \cup \Theta_3 \cup \Theta_6$, $\Theta_U = \Theta_4 \cup \Theta_5 \cup \Theta_7$, $\Theta_C = \Theta_8$, and $\Phi = \Phi_1 \cup \Phi_2$.

Fixed parameters: The parameters of the four transformations f_1 , f_2 , f_3 , and f_6 are fixed to 1. For f_1 , f_2 , or f_3 , θ^c is fixed to 1 to erase the high-frequency



Fig. 5: Comparison of obfuscated images with or without the BRS optimization.

information or image details with a strong low-pass filter. In contrast, for f_7 , the amplitude of the checkerboard pattern is fixed to $0.3 \times \theta^c = 0.3$ experimentally to mimic the high-frequency information in the original image and achieve MD.

Uniform parameters: For the three transformations f_4 , f_5 , and f_7 , the parameters are initialized uniformly. Note that, in Table 1, θ -range specifies the range $[\theta_{\min}, \theta_{\max}]$ of each parameter θ in f_n . We initialize θ according to the uniform distribution $\mathcal{U}(\theta_{\min}, \theta_{\max})$ with the probability density function, given by

$$p(\theta) = \begin{cases} \frac{1}{\theta_{\max} - \theta_{\min}} & \text{if } \theta_{\min} \leq \theta \leq \theta_{\max}, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Color parameters: For *Color Scaling* (f_8), an input color is multiplied by a parameter θ to yield an output color in (3). If $\theta > 1$, it is a brightening function. On the contrary, if $0 < \theta < 1$, it is a darkening one. Hence, we first decide with the equal probabilities of $\frac{1}{2}$ whether to choose a brightening or darkening function. Then, for a brightening one, the scaling parameter θ is initialized according to the uniform distribution $\mathcal{U}(1, \theta_{\max})$. On the other hand, for a darkening one, the inverse $\frac{1}{\theta}$ is initialized according to $\mathcal{U}(1, \theta_{\max})$. Thus, θ is determined within $[\frac{1}{\theta_{\max}}, \theta_{\max}]$. As specified in Table 1, $\theta_{\max} = \frac{11}{10}$.

Composing parameters: Each composing parameter ϕ for superposing the results of f_1 , f_2 , and f_3 in (4) or f_5 , f_6 , and f_7 in (5) is initialized according to the uniform distribution $\mathcal{U}(0, 1)$. Note that these composing parameters are normalized to $\bar{\phi}$ using the softmax operator for the convex combination.

3.4 Parameter Optimization

A transformed image $I_{\text{out}}(\Theta)$, which is obtained by applying the composite transformation with randomly initialized parameters in Θ to an input image I_{in} , may be indecipherable by a machine. More specifically, when it is processed by a feature extractor, its output vector $\mathcal{F}(I_{\text{out}}(\Theta))$ may deviate from the original vector $\mathcal{F}(I_{\text{in}})$ excessively, yielding a wrong decision result, *e.g.* a wrong identity. Hence, to improve the MD, we adopt the BRS technique [14] to optimize the parameters in Θ so that the distance $\|\mathcal{F}(I_{\text{out}}(\Theta)) - \mathcal{F}(I_{\text{in}})\|_2$ between the two feature vectors is minimized and their cosine similarity $\frac{\mathcal{F}(I_{\text{out}}(\Theta))^T \mathcal{F}(I_{\text{in}})}{\|\mathcal{F}(I_{\text{out}}(\Theta))\| \|\mathcal{F}(I_{\text{in}})\|}$ is maximized. Fig. 5 shows examples of the BRS optimization. We see that the optimization does not change the level of HI noticeably. However, it improves the degree of MD effectively, as will be shown in Section 4.4.

BRS [14] was originally proposed for interactive image segmentation to optimize an interaction map to refine a segmentation result. In this work, we adopt BRS to optimize the parameters in Θ of the composite transformation. To this end, we define an energy function $\mathcal{E}(\Theta)$ appropriate for face obfuscation and develop its minimization scheme.

For uniform parameters, we have two types of θ -ranges in Table 1: $[\theta_{\min}, \theta_{\max}]$ or $[-\theta_{\max}, \theta_{\max}]$. In both types, the obfuscating functions are designed so that HI increases, or at least maintains the same level, as $|\theta|$ gets closer to $\theta_{\max}$. Thus, to improve HI, we define the HI energy for uniform parameters as

$$\mathcal{E}_U = \sum_{\theta_i \in \Theta_i \subset \Theta_U} \max\{\lambda_i - |\theta_i|, 0\} \quad (8)$$

where λ_i is the margin for Θ_i specified in Table 1. In BRS, the partial derivative of $\mathcal{E}_U$ with respect to each uniform parameter θ_i is required, which is given by

$$\frac{\partial \mathcal{E}_U}{\partial \theta_i} = \begin{cases} -1 & \text{if } 0 < \theta_i < \lambda_i, \\ 1 & \text{if } -\lambda_i < \theta_i < 0, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Similarly, we define the HI energy for color parameters as

$$\mathcal{E}_C = \sum_{\theta_i \in \Theta_i \subset \Theta_C} \max\{[\theta_i > 1](\lambda_i - \theta_i), [\theta_i < 1](\theta_i - \frac{1}{\lambda_i})\} \quad (10)$$

where $[\cdot]$ is an indicator function. Note that, if $\theta_i = 1$, the corresponding color is not scaled in f_s . Hence, by minimizing $\mathcal{E}_C$, it is encouraged that θ_i is as different from 1 as possible. The derivative of $\mathcal{E}_C$ is given by

$$\frac{\partial \mathcal{E}_C}{\partial \theta_i} = \begin{cases} -1 & \text{if } 1 < \theta_i < \lambda_i, \\ 1 & \text{if } \frac{1}{\lambda_i} < \theta_i < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

On the other hand, to improve MD, the distance between the feature vectors should be reduced. Hence, we define the MD energy of the distance $\mathcal{E}_D$ and the MD energy of the cosine similarity $\mathcal{E}_S$ as

$$\mathcal{E}_D = \|\mathcal{F}(I_{\text{out}}(\Theta)) - \mathcal{F}(I_{\text{in}})\|_2, \quad (12)$$

$$\mathcal{E}_S = 1 - \frac{\mathcal{F}(I_{\text{out}}(\Theta))^T \mathcal{F}(I_{\text{in}})}{\|\mathcal{F}(I_{\text{out}}(\Theta))\| \|\mathcal{F}(I_{\text{in}})\|}. \quad (13)$$

Then, we adopt the overall energy function given by

$$\mathcal{E} = \mathcal{E}_U + \mathcal{E}_C + \mathcal{E}_D + \mathcal{E}_S \quad (14)$$

whose derivative is

$$\frac{\partial \mathcal{E}}{\partial \Theta} = \frac{\partial \mathcal{E}_U}{\partial \Theta} + \frac{\partial \mathcal{E}_C}{\partial \Theta} + \frac{\partial \mathcal{E}_D}{\partial \Theta} + \frac{\partial \mathcal{E}_S}{\partial \Theta}. \quad (15)$$

Table 2: Quantitative comparison of face obfuscation results. In each test, the best result is **boldfaced**, while the second best is underlined.

Dataset $\triangleright$	Human	Face Verification		Age	Gender	Attribute
	LFW	LFW	IJB-C	Adience	Adience	CelebAMask-HQ
Metric $\triangleright$	ACR ($\downarrow$)	Acc. ($\uparrow$)	Acc. ($\uparrow$)	Acc. ($\uparrow$)	Acc. ($\uparrow$)	Acc. ($\uparrow$)
Original	–	99.82	94.64	65.24	85.65	88.02
Blur	0.8378	79.20	81.82	43.33	31.36	61.31
Pixelate	0.7622	85.58	83.54	22.60	64.97	62.74
FALCO [1]	0.6844	78.68	71.51	20.24	71.53	72.88
PRO-Face (Blur) [43]	0.8978	<u>98.10</u>	87.63	<u>57.01</u>	40.86	65.04
PRO-Face (Pixelate) [43]	0.8133	94.35	86.93	23.66	72.18	66.79
PRO-Face (FaceShifter) [43]	0.8333	96.72	80.23	21.14	75.68	75.13
PRO-Face (SimSwap) [43]	0.7800	96.13	78.43	21.51	65.43	75.47
Forbes-G	0.2578	96.10	89.47	49.82	<u>78.64</u>	<u>73.18</u>
Forbes-T	0.2422	98.72	90.89	59.03	81.26	76.20

The first and second terms can be computed using (9) and (11), respectively, while the last two terms $\frac{\partial \mathcal{E}_D}{\partial \Theta}$ and $\frac{\partial \mathcal{E}_S}{\partial \Theta}$ are computed by BRS. Thus, we minimize $\mathcal{E}$ using the L-BFGS method [26] and determine an optimal set of parameters Θ^* . However, the fixed parameters are not modified in the optimization step, so $\Theta_F^* = \Theta_F$.

3.5 Implementation Details

It is worth pointing out that the proposed Forbes algorithm operates at the inference stage, so it does not need any training. Forbes divides a facial region into 7×7 blocks. In other words, we set $M = N = 7$. We analyze the impacts of M and N on the obfuscating performance in Section 4.4.

Forbes aims to generate an obfuscated image that can be deciphered by machines only, so we may use several pre-trained networks in parallel as the feature extractor in Fig. 3. However, the inference time of Forbes depends on the feature extractor, since Forbes should compute the derivative of the energy function. If we combine more pre-trained networks in parallel, the obfuscated image may retain more information decipherable by machines, but the parameter optimization becomes slower in general. Therefore, to reduce the inference time, we may use a task-specific feature extractor, instead of several networks. Thus, we implement two types of Forbes: Forbes-G and Forbes-T.

Forbes-G: In Forbes-G, we employ five pre-trained networks in parallel as the feature extractor regardless of the task. Specifically, we use pre-trained AdaFace [16], GOL [20], SphereNet [13], MobileNetV2 [34], and ResNet-50 [10] as the feature extractor. Thus, the output image generated by Forbes-G tends to contain identity, age, gender, and facial attributes.

Forbes-T: Depending on the required information for a given task, we employ a task-specific feature extractor in Forbes-T. For example, we adopt the feature

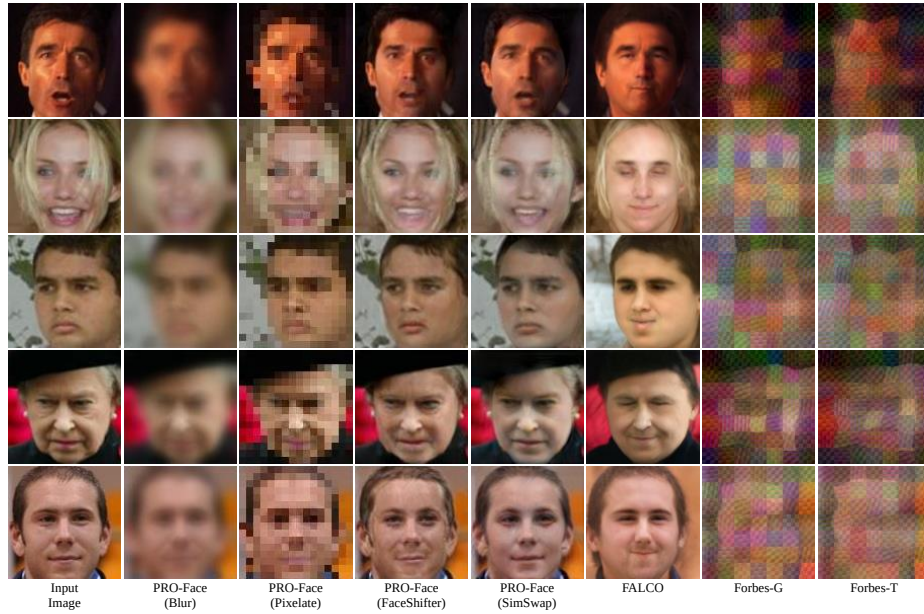


Fig. 6: Qualitative comparison of obfuscated images in the LFW dataset.

extractor of a face verification network to retain the identity information or that of a gender classification network to retain the gender information.

4 Experiments

We assess Forbes on various benchmark datasets for face-related vision tasks.

4.1 Datasets and Evaluation Protocols

Face verification: LFW [11] consists of 13,233 facial images with 5,749 identities. We use the test set containing 6,000 pairs of facial images with half matched and half unmatched. IJB-C [29] consists of 138,000 facial images. We randomly sample 20,000 pairs with half matched and half unmatched. To use the existing face verification network [16], we align 112×112 facial regions using landmarks detected by MTCNN [44]. We measure the verification accuracy of AdaFace [16] using their official model. We verify whether two facial images, reference I_{ref} and source I_{out} , belong to the same identity. Note that I_{out} is obfuscated from an input image I_{in} that is different from I_{ref} . This face verification tests demonstrate the identity-preserving capability of Forbes.

Age estimation & gender classification: For these two tasks, we use the Adience dataset [8], composed of 26,580 facial images of 2,284 identities. For age estimation, we measure the accuracy score using the GOL algorithm [20],

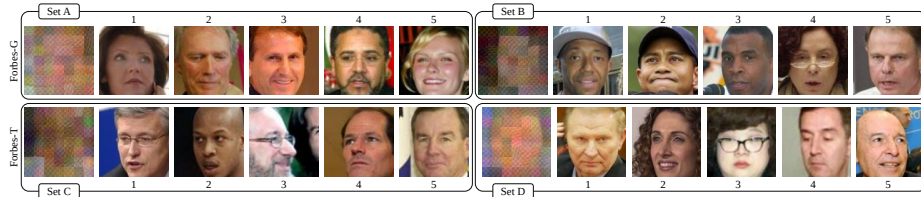


Fig. 7: Examples of the questions in the user study. A subject was required to select one image, from the five candidates, containing the same identity as the obfuscated image. The correct answers are A) 3, B) 4, C) 2, D) 1.

which performs the k-NN estimation in a feature space. We train GOL with the provided settings and conduct the 5-fold cross-validation on Adience [8,36]. We use the official model of SphereNet [13] for gender classification.

Facial attribute classification: CelebAMask-HQ [17] contains 30,000 celebrity images selected from CelebA [27]. Each face of resolution 1024×1024 is annotated with 40 attribute labels (*e.g.* eyeglasses, hair color, and mustache). We test the capability of retaining facial attributes after obfuscation. For attribute classification, we train MobileNetV2 [34] on CelebA.

4.2 Comparative Results

Table 2 compares the accuracies of Forbes on face verification, age estimation, gender classification, and facial attribute classification with those of previous methods [1,43]. Note that Forbes-T significantly outperforms the existing algorithms on all the tasks. In particular, compared to the original images, Forbes-T degrades the face verification accuracies only marginally, since it successfully retains identity information decipherable by machines. Moreover, even Forbes-G provides better results on the tasks of face verification on IJB-C, gender classification, and facial attribute classification than the existing algorithms.

Fig. 6 compares obfuscated images in the LFW dataset qualitatively. We see that PRO-Face (Blur) retains some facial information recognizable by humans, although it achieves a higher accuracy on the LFW dataset than Forbes-G does. Similarly, the other versions of PRO-Face fail to achieve sufficient HI. Also, FALCO [1], which is a face anonymization method, retains some facial attributes, lowering the HI. In contrast, both Forbes-G and Forbes-T make it challenging for humans to decipher the obfuscated images. More qualitative results on the other datasets are presented in Fig. S-1 in the supplement.

4.3 User Study

Forbes improves HI by minimizing the HI energies in (8) and (10). To assess the HI of obfuscated images, we conducted a user study. We compared Forbes with the previous methods. Specifically, we used Blur, Pixelate, PRO-Face (Blur), PRO-Face (Pixelate), PRO-Face (FaceShifter), PRO-Face (SimSwap), FALCO,

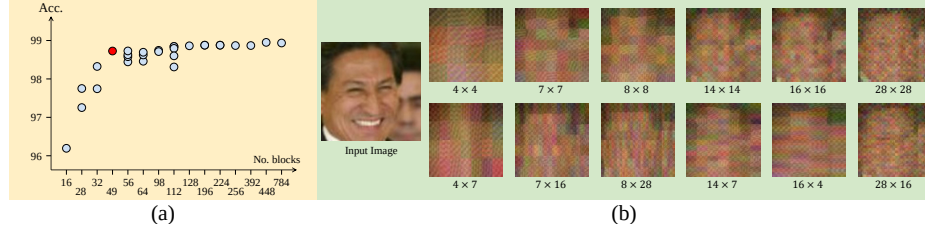


Fig. 8: Impacts of the number $M \times N$ of blocks on obfuscation results. (a) Verification accuracies of Forbes-T on the LFW dataset according to $M \times N$. (b) Qualitative results.

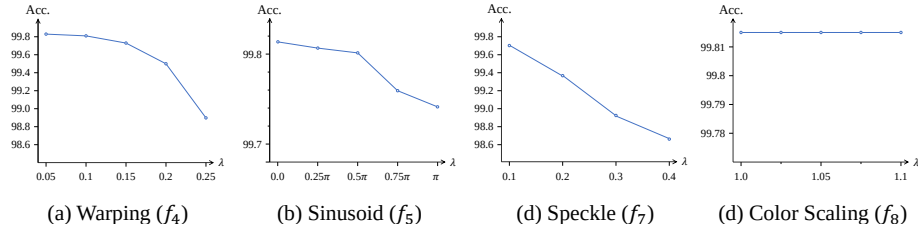


Fig. 9: Verification accuracies of single obfuscating transformations f_4 , f_5 , f_7 , and f_8 according to margins λ .

Forbes-G, and Forbes-T to generate 30 obfuscated images, respectively. Each obfuscated image was shown to a subject with five reference images: one with the same identity and the other four with different identities. The input image for generating the obfuscated image was different from the reference image of the same identity; they did not share the same background and poses.

We recruited 15 volunteers, who were asked to choose one reference image, which had the same identity as the obfuscated image, from the five candidates. We defined the correctness rate as the ratio of the right answers over the 30 tests. Then, we computed the average correctness rate (ACR) of the 15 volunteers.

Table 2 compares these user study results on the LFW dataset. Since we provided multiple choice problems of five reference images, the ACR of random guess should be around 0.2. Note that, compared with Blur and Pixelate, the two PRO-Face versions worsen the ACR scores. In contrast, the proposed Forbes-T yields a low ACR of 0.2422, close to 0.2. This indicates that Forbes achieves high HI. Fig. 7 shows examples of the questions in this user study. More examples are presented in Fig. S-3 and S-5 in the supplement.

4.4 Analysis

We conduct ablation studies to analyze the main components of Forbes: block size, margins for the HI energies, energy functions, and obfuscating transformations.

Block size: We analyze the impacts of the number $M \times N$ of blocks on obfuscation results in Fig. 8. We test many block configurations from 4×4 to

Table 3: Ablation studies for the energy functions on the LFW dataset [11]. Verification accuracies of AdaFace [16] are reported.

Method	$\mathcal{E}_U$	$\mathcal{E}_C$	$\mathcal{E}_D$	$\mathcal{E}_S$	Forbes-T
I					56.30
II	✓				54.49
III		✓			54.64
IV			✓		98.90
V				✓	99.11
VI	✓	✓			52.18
VII	✓		✓		81.42
VIII	✓			✓	87.67
IX		✓	✓		83.30
X		✓		✓	89.98
XI			✓	✓	99.27
XII	✓	✓	✓		83.14
XIII	✓	✓		✓	83.10
XIV	✓		✓	✓	93.12
XV		✓	✓	✓	96.68
XVI	✓	✓	✓	✓	98.70

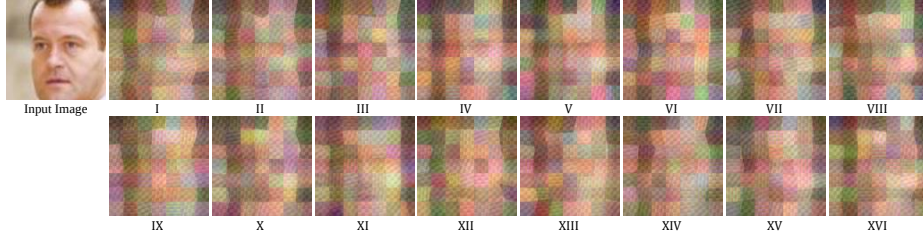


Fig. 10: Examples of obfuscated images by the methods in Table 3.

28×28 and plot the verification accuracies of Forbes-T according to the number of blocks. In the graph, as more blocks are used, the verification accuracy improves in general. However, we see that, with more blocks, visually salient information remains in the obfuscated image, lowering HI. Hence, as a tradeoff between HI and MD, we use 7×7 blocks in the Forbes algorithm.

Margins: We also analyze the impacts of the margins λ for the HI energies in (8) and (10). Fig. 9 plots the verification accuracies of single obfuscating transformations f_4 , f_5 , f_7 , and f_8 according to margin values. We see that, for f_4 , f_5 , and f_7 , a larger margin corresponds to a lower machine accuracy. In other words, for MD, it is desirable to set a small margin. However, except f_5 , the HI energies are designed to achieve a higher HI level with a larger margin. Hence, for HI, it is better to set a large margin. Note that, in f_5 , the axis angles of the sinusoidal noise are used as parameters, which have no impact on HI. Therefore, by considering both MD and HI, we select margins 0.05, 0.0, 0.1, and 1.05 for f_4 , f_5 , f_7 , and f_8 , respectively, as listed in Table 1.

Energy functions: Table 3 lists the verification accuracies on the LFW dataset according to combinations of the energy functions in the parameter opti-

Table 4: Ablation studies for the transformations on the LFW dataset [11]. Verification accuracies of AdaFace [16] are reported.

Method	Averaging	Warping	Noising	Scaling	Forbes-T
I	✓				96.13
II	✓	✓			96.91
III	✓	✓	✓		98.12
IV	✓	✓	✓	✓	98.70



Fig. 11: Examples of obfuscated images by the methods in Table 4.

mization. Fig. 10 shows qualitative results of these combinations. First, method I does not use any energy functions; it uses randomly initialized parameters without optimization. Therefore, it yields the lowest accuracy. Without the HI energies $\mathcal{E}_U$ and $\mathcal{E}_C$, methods IV, V, and XI can focus on MD only and thus yield slightly better accuracies than XVI. However, in Fig. 10, they yield lower levels of HI than XVI. On the contrary, by comparing XVI with the methods excluding the MD energies $\mathcal{E}_D$ and $\mathcal{E}_S$, we see that the verification accuracies degrade severely without the MD energies. In Fig. 10, method I and the methods without the MD energies have similar HI levels. Using all four energies, the proposed Forbes (XVI) achieves a good balance between HI and MD.

Obfuscating transformations: We fix the parameter values of f_1, f_2, f_3 , and f_6 to 1 to yield a high HI level. In Table 4 and Fig. 11, method I effectively erases the high-frequency information while achieving a machine accuracy of 96.13%. Method II, which applies f_4 to the superposed image, warps the image and partly reconstructs high-frequency information. Note that method II yields a slightly better accuracy than method I. Since the warping cannot recover the information completely, method III deceives the machine by adding a noise map to the warped image. Moreover, we adopt f_8 to scale each block and make it easy for machines to decipher, yielding the highest MD.

5 Conclusions

We proposed a novel algorithm for face obfuscation, called Forbes, to strike a good balance between HI and MD. To this end, we designed eight parameterized obfuscating transformations, combined them into a composite transformation, and developed an optimization scheme for the parameters based on BRS. Through extensive experiments on various facial datasets, it was demonstrated that the proposed Forbes achieves face obfuscation more effectively than the conventional algorithms.

Acknowledgements

This work was supported by the 42dot Inc. and the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B03002310 and No. RS-2024-00397293).

References

1. Barattin, S., Tzelepis, C., Patras, I., Sebe, N.: Attribute-preserving face dataset anonymization via latent code optimization. In: CVPR (2023)
2. Brkić, K., Sikirić, I., Hrkać, T., Kalafatić, Z.: I Know That Person: generative full body and face de-identification of people in images. In: CVPRW (2017)
3. Chen, J., Konrad, J., Ishwar, P.: VGAN-based image representation learning for privacy-preserving facial expression recognition. In: CVPRW (2018)
4. Cho, D., Lee, J.H., Suh, I.H.: CLEANIR: controllable attribute-preserving natural identity remover. *Applied Sciences* **10**(3), 1120 (2020)
5. Ciftci, U.A., Yuksek, G., Demir, I.: My Face My Choice: privacy enhancing deep-fakes for social media anonymization. In: WACV (2023)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: CVPR (2019)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(2), 295–307 (2015)
8. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Secur.* **9**(12), 2170–2179 (2014)
9. Hassan, E.T., Hasan, R., Shaffer, P., Crandall, D., Kapadia, A.: Cartooning for enhanced privacy in lifelogging and streaming videos. In: CVPRW (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition (2008)
12. Hukkelås, H., Mester, R., Lindseth, F.: DeepPrivacy: A generative adversarial network for face anonymization. In: ISVC (2019)
13. Hung, C.Y., Tu, C.H., Wu, C.E., Chen, C.H., Chan, Y.M., Chen, C.S.: Compacting, picking and growing for unforgetting continual learning. In: NeurIPS (2019)
14. Jang, W.D., Kim, C.S.: Interactive image segmentation via backpropagating refinement scheme. In: CVPR (2019)
15. Jun, J., Lee, J.H., Lee, C., Kim, C.S.: Depth map decomposition for monocular depth estimation. In: ECCV (2022)
16. Kim, M., Jain, A.K., Liu, X.: AdaFace: quality adaptive margin for face recognition. In: CVPR (2022)
17. Lee, C.H., Liu, Z., Wu, L., Luo, P.: MaskGAN: towards diverse and interactive facial image manipulation. In: CVPR (2020)
18. Lee, S.H., Kim, C.S.: Deep repulsive clustering of ordered data based on order-identity decomposition. In: ICLR (2020)
19. Lee, S.H., Kim, C.S.: Order learning using partially ordered data via chainization. In: ECCV (2022)
20. Lee, S.H., Shin, N.H., Kim, C.S.: Geometric order learning for rank estimation. In: NeurIPS (2022)
21. Lee, S.H., Shin, N.H., Kim, C.S.: Unsupervised order learning. In: ICLR (2024)
22. Li, D., Shi, X., Zhang, Y., Cheung, K.C., See, S., Wang, X., Qin, H., Li, H.: A simple baseline for video restoration with grouped spatial-temporal shift. In: CVPR (2023)
23. Li, J., Han, L., Chen, R., Zhang, H., Han, B., Wang, L., Cao, X.: Identity-preserving face anonymization via adaptively facial attributes obfuscation. In: ACM MM (2021)

24. Li, J., Han, L., Zhang, H., Han, X., Ge, J., Cao, X.: Learning disentangled representations for identity preserving surveillance face camouflage. In: ICPR (2021)
25. Li, J., Zhang, H., Liang, S., Dai, P., Cao, X.: Privacy-enhancing face obfuscation guided by semantic-aware attribution maps. *IEEE Trans. Inf. Forensics Secur.* **18**, 3632–3646 (2023)
26. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. *Mathematical programming* **45**(1-3), 503–528 (1989)
27. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
28. Maximov, M., Elezi, I., Leal-Taixé, L.: CIAGAN: conditional identity anonymization generative adversarial networks. In: CVPR (2020)
29. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., Grother, P.: Iarpa janus benchmark - c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB) (2018)
30. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: MagFace: a universal representation for face recognition and quality assessment. In: CVPR (2021)
31. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
32. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: ECCV (2018)
33. Rosberg, F., Aksoy, E.E., Englund, C., Alonso-Fernandez, F.: FIVA: facial image and video anonymization and anonymization defense. In: ICCVW (2023)
34. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: CVPR (2018)
35. Sarriidis, I., Koutlis, C., Papadopoulos, S., Diou, C.: FLAC: fairness-aware representation learning by suppressing attribute-class associations. *arXiv preprint arXiv:2304.14252* (2023)
36. Shin, N.H., Lee, S.H., Kim, C.S.: Moving Window Regression: a novel approach to ordinal regression. In: CVPR (2022)
37. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: CVPR (2018)
38. Sun, Q., Ma, L., Oh, S.J., Van Gool, L., Schiele, B., Fritz, M.: Natural and effective obfuscation by head inpainting. In: CVPR (2018)
39. Vishwamitra, N., Knijnenburg, B., Hu, H., Kelly Caine, Y.P., et al.: Blur vs. Block: investigating the effectiveness of privacy-enhancing obfuscation for images. In: CVPRW (2017)
40. Voigt, P., Von dem Bussche, A.: The EU general data protection regulation (GDPR). A Practical Guide, 1st Ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)
41. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. *Int. J. Comput. Vis.* **127**, 1106–1125 (2019)
42. Yasarla, R., Perazzi, F., Patel, V.M.: Deblurring face images using uncertainty guided multi-stream semantic networks. *IEEE Trans. Image Process.* **29**, 6251–6263 (2020)
43. Yuan, L., Liu, L., Pu, X., Li, Z., Li, H., Gao, X.: PRO-Face: a generic framework for privacy-preserving recognizable obfuscation of face images. In: ACMMM (2022)
44. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* **23**(10), 1499–1503 (2016)

45. Zhao, Z., Liu, Q., Zhou, F.: Robust lightweight facial expression recognition network with label distribution training. In: AAAI (2021)