Facial Image Privacy Protection Based on Principal Components of Adversarial Segmented Image Blocks

JINGJING YANG¹23, JIAXING LIU², AND JINZHAO WU¹345
¹Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, China
²School of Information Science and Engineering, Hebei North University, Zhangjiakou 075000, China
³University of Chinese Academy of Sciences, Beijing 100049, China
⁴Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi University for Nationalities, Nanning 530006, China
⁵School of Computer Science and Electronic Information, Guangxi University, Nanning 530000, China

Corresponding author: Jinzhao Wu (yjr78z@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772006, in part by the Science and Technology Major Project of Guangxi under Grant AA17204096, in part by the Key Research and Development Project of Guangxi under Grant AB17129012, in part by the Fund Project of Hebei North University under Grant JYT2019016, the Financial Supported Projects of Zhangjiakou Science and Technology under Grant 1911016C-9, in part by the Natural Science Foundation of Hebei Province of China, and in part by the Special Fund for Bagui Scholars of Guangxi.

ABSTRACT The features in facial images, which are utilized for a variety of technological applications, pose a significant privacy concern for users. This paper proposes a method for protecting privacy in facial images based on the principal components of adversarial segmented image blocks. Generative adversarial network parameters are compressed by segmenting the facial images into blocks and extracting the principal components of the segmented image. The generator and discriminator in the generative adversarial network then generate images similar to the original facial images; the facial images generated by the generator, as-driven by the target recognition network, markedly different from the original facial images. As the generator, discriminator, and target recognition network compete with each other, minor perturbation is added to the principal components of the facial images to protect the users’ privacy and prevent distinct face-related features of the images from being easily extracted. Experimental results show that the proposed method outperforms other similar methods in terms of generated image quality, operation speed, and target recognition network accuracy.

INDEX TERMS Facial image privacy protection, generative adversarial network, principal components, adversarial samples.

I. INTRODUCTION Modern facial image recognition technology is relatively unaffected by problems with lighting [1], [2], or occlusion [3], [4]. It has been widely applied in the Internet of Things, security, mobile payment, and many other applications. Passenger vehicles, for example, use face recognition for keyless starting [5] – merchants use face recognition for online payment [6], and buildings use face recognition for access control [7]. Facial features come with significant privacy concerns [8]. According to incomplete statistics, the number of “selfie” photos shared daily by users on social media now exceeds one billion [9], [10]. Most social networks impose no limits on the downloading of facial images. To this effect, the leakage of facial images poses a significant threat. If features in the facial images are extracted by lawbreakers and data mining is performed to obtain user-related privacy information, the users’ property may become severely insecure [11]–[13]. Protecting the facial image features can effectively preserve user privacy. It is necessary to secure these features from extraction before allowing it to be published while retaining as much of the original information as possible to ensure the readability and practicality of the image.

There are advantages and drawbacks to the existing methods for facial image privacy protection. Methods based on
encryption [14], [15], for instance, require burdensome calculations, have poor real-time performance, do not allow for the direct use of data, and are restricted within an application scope. Methods based on image filtering [16] significantly damage the original face information and do not guarantee image availability. Anonymity-based methods [17], [18] are susceptible to privacy leakage due to similar or background knowledge attacks.

In 2013, Szegedy et al. [19] proposed the “adversarial samples” concept wherein a small amount of perturbation is added to sample data to cause the target model to generate incorrect classification with high confidence. Wide range protection of similar image privacy protection methods then emerged, including the Fast Gradient Sign Method (FGSM) [20], Adversarial Patch [21], and One Pixel Attack [22]. Xiao et al. [23] proposed the generation of adversarial examples with adversarial networks (AdvGAN) based on generative adversarial networks in 2018. Adding a small amount of perturbation to the image with a generative adversarial network (GAN) causes the target network to consistently produce the wrong classification and is robust against both white- and black-box attacks. He et al. [24] and Wu et al. [25] proposed networks similar to AdvGAN for the generation of facial privacy images, where the target network is misled to perform a misclassification while ensuring the availability of facial images.

The paper proposes a method for facial image privacy protection based on principal components of adversarial segmented image blocks. We add tiny perturbations to the facial images. When the facial images are published on the cloud service platform such as social media, the recognition network of potential lawbreakers will recognize errors. As a result, the data mining of users by potential criminals will not be successful while protecting the users’ privacy and maintaining the practicability of pictures. The main contributions are as follows:

1. A tiny amount of perturbation is added to the principal components of facial images to ensure that the image as-generated is fully available.
2. The principal components of the segmented facial images are extracted to minimize the generator input parameters and accelerate the running speed of the process.
3. A target face recognition network is added in the competition between the generator and the discriminator for misleading, which gives the generated facial image a different label than the original image.
4. Peak Signal to Noise Ratio (PSNR) constraint conditions is added to ensure the generated facial image is similar to the original facial image pixel-wise.

II. RELATED WORK

Goodfellow et al. [20] proposed a fast gradient sign method (FGSM) in 2014. The gradient of the input image is obtained by calculating the target category, and then the sign function of the gradient is obtained. Finally, the obtained results are added to the original image as the antagonistic perturbations to obtain the antagonistic samples that can confuse the recognition network. Liu et al. [35] in 2018 and Linardos et al. [36] in 2019 applied FGSM in face privacy protection. Both can protect the privacy of images while maintaining high quality. However, FGSM has the disadvantage that the added perturbations can be easily removed, such as using the median filtering method. So FGSM is not in our consideration.

Influenced by AdvGAN proposed by Xiao et al. [23], He et al. [24] proposed a picture privacy protection algorithm based on the generated adversary network (PriGAN) in 2019. The U-NET network structure is combined with the generated adversarial network (GAN) to realize image privacy protection. However, the resulting privacy images have a checkerboard effect. In 2019, Wu et al. [25] proposed a new architecture for image Privacy protection named Privacy-Protective-GAN (PP-GAN), adding validators and adjustment modules explicitly designed for face recognition to achieve de-identified output with a similar structure based on a single input. However, the features of the generated privacy image are different from the original image, so it is not suitable for some application scenarios that require high practicability of the image. We will propose a method that is superior to other similar methods in terms of image quality, operation speed, and target recognition network accuracy.

III. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis [26] (PCA) is often used to reduce the dimensionality of data and extract the main features of images. PCA works based on relatively little operating data and with a relatively small memory footprint, which makes it accessible in many fields, including image classification.

A. BACKGROUND

N facial images are given \( X = \{X_1, X_2, \ldots, X_N\} \) for any facial image \( X_i \in \mathbb{R}^{m \times n} \). Singular Value Decomposition (SVD) for PCA dimensionality reduction is applied to \( X \), as follows:

\[
X_{i,m,n} = U_{m,m}A_{m,n}V_{n,n}^T \approx U_{m,k}A_{k,k}V_{k,n}^T \tag{1}
\]

where \( U_{m,m} \) is the left singular matrix of \( X_i \), which can compress the \( X_i \) number of rows, \( V_{n,n} \) is the right singular matrix of \( X_i \), which can compress the \( X_i \) number of columns, and \( A_{m,n} \) is the singular value of \( X_i \) that has been sorted by size. As the sum of the first \( k \) singular values accounts for more than 95% of the sum of all singular values, the left singular matrix of the first \( k \) dimensions can be multiplied by the singular value and right singular matrix, which is approximately equal to the original image \( X_i \). This results in image dimensionality reduction.

B. PERTURBATION OF PRINCIPAL COMPONENTS

For convenience, let the compression matrix be \( Z = AV^T \). Then, (2) holds for any facial image \( X \).

\[
X_i = U_{m,m}Z_{m,n} \approx U_{m,k}Z_{k,n} \tag{2}
\]
Xᵢ can be regarded as a linear transformation of matrices U and Z. The multiplication of matrices U and Z is actually an asymmetric transformation of the elements in Z and the transformation of a single element Z into a single element in Xᵢ as marked with a solid red line in Figure 1. As shown in (3), the elements in the matrix U may be subjected to slight perturbation; the results of the asymmetric transformation then also generate corresponding fluctuations in various directions, as shown in Figure 1. Possible fluctuations in the transformation of a single element in Z (as marked with a dotted line in the figure) also cause the corresponding pixel point in the facial image Xᵢ to change slightly.

\[ U' = U + N \]  
(3)

where the matrix N is a perturbation matrix with the same dimensions as U.

C. PERTURBATION ANALYSIS

The calculated adversarial sample of U was actually slight perturbation with the base coordinates of the principal components. The elements in U have been standardized, and the value range is within \([-1, 1]\), which contributes to the iterative operation of the algorithm for perturbation calculation. The principal component contains the main features of the facial image Xᵢ. Accumulating perturbation to the principal component via linear transformation can speed up the generation of facial images with privacy protected. As the principal components contain the main features of the facial image Xᵢ, when the perturbation is added to the main features, the perturbation in the generated facial image with privacy protected cannot be easily filtered out. That is, the image cannot be quickly restored.

For the generated set of facial images with protected privacy \(X' = \{X'_1, X'_2, \ldots, X'_p\}\), in the case of \(X'_i \in X'\), (4) is true.

\[ X'_i = U'Z + E = X_i + NZ + E \]  
(4)

where E is the error between \(X_i\) that was subjected to PCA transformation and \(X'_i\). To ensure the availability of the facial privacy image, the difference between the facial image \(X_i\) and the privacy image \(X'_i\) should be invisible to the naked eye; that is, the error and perturbation matrix should be minimal.

Let \(f\) be the fitting function of the target face recognition network. To protect the privacy of the facial image for \(f\), the difference between the facial image \(X_i\) and the privacy image \(X'_i\) should be maximized, as shown in (5).

\[ \min ||X_i - X'_i||_2 \quad \text{s.t.} \quad f(X_i) \neq f(X'_i) \]  
(5)

D. IMAGE SEGMENTATION

The facial image \(X_i\) usually has high dimensionality. Here, \(X_i\) was segmented into sub-blocks of \(p \times q\) as follows:

\[ X_i = \begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & \ddots & \vdots \\ X_{p1} & \cdots & X_{pq} \end{pmatrix} \]  
(6)

Establishing the facial privacy image \(X'_i\) is equivalent to adding perturbation to the left singular matrix U of the sub-block \(X_i\), as shown in (7). As the left singular matrix \(U\) of the sub-block is much smaller in dimensions than the left singular matrix \(U\) of the entire facial image \(X_i\), the number of parameters in the network can be reduced. The sub-blocks obtained by segmentation also increase the number of trained samples and decrease the complexity of the problem, thus accelerating the perturbation-solving process.

\[ X'_i = \begin{pmatrix} (U_{11} + N_{11})Z_{11} & \cdots & (U_{1q} + N_{1q})Z_{1q} \\ \vdots & \ddots & \vdots \\ (U_{p1} + N_{p1})Z_{p1} & \cdots & (U_{pq} + N_{pq})Z_{pq} \end{pmatrix} \]  
(7)

IV. MATH FACIAL IMAGE GENERATION NETWORK OF ADVERSARIAL SEGMENTED IMAGE BLOCK PRINCIPAL COMPONENTS

As the data distribution of the facial privacy image \(X'_i\) is infinitely approximated to the original facial image \(X_i\), a GAN can be applied to solve the facial privacy image that satisfies (5).

A. BACKGROUND

The GAN is a deep learning model proposed by Goodfellow et al. [27] in 2014. The loss function of \(G\) and \(D\) is minimized by finding the Nash equilibrium point between the generator \(G\) and the discriminator \(D\), so that the data generated from \(G\) approximates the real data.

Goodfellow et al. [27] first proposed the Conditional Generative Adversarial Net (CGAN), where parameters guide the generation of data under supervised learning. GAN techniques proposed by Wang et al. [29], Xiao et al. [23], He et al. [24], and Wu et al. [25] are improved variations of CGAN. The facial image generation network of the adversarial segmentation principle components can also be considered a supervised learning process. This differs from the CGAN process as a supervised network is added to the generative adversarial network to drive classification errors in the generated facial images.
the parameters represents the original image and \( num = p \times q \) represents the number of image blocks. The label in the original matrix is retained after segmentation and represents the set of block images in the image merging function \( \Gamma(x, num) \). The matrix is restored according to the label of the block matrix in the original image.

### C. LOSS FUNCTION

The role of the discriminator \( D \) is to distinguish the difference between \( X' \) and the original facial image \( X \), thus ensuring close consistency in data distributions between \( X' \) and the original facial image \( X \). The first loss function of PcadvGAN is expressed as follows:

\[
L_{\text{GAN}} = \mathbb{E}_X \log D(X) + \mathbb{E}_{X',X'} \log(1 - D(\Gamma'(X')|G(\beta(X))))
\]

(8)

In (8), the original facial image \( X \) participates in the generation of the facial privacy image \( X' \), which is essentially a set \( N \) to which adversarial sample perturbation generated by the generator is added. The \( X' \) with the added adversarial sample perturbation then trains the discriminator \( D \) together with the original facial image \( X \). The generator \( G \) is reversely updated, according to the probability that the outputs of the samples by the discriminator \( D \) are true, until the loss function \( L_{\text{GAN}} \) reveals the optimal value. The error \( E \) between \( X \) subjected to PCA transformation and \( X' \) as the sample perturbation set \( N \) is thus minimized.

The target face recognition network \( f \) is denoted as a network model that has been properly trained using the dataset at high accuracy. The parameters would not be updated during the training process. The proposed method (Table 1) can thus be applied to combat black-box and white-box attacks on the samples [30]. As shown in (9), \( f \) receives a privacy image \( X' \) as an input. If it belongs to the classification \( t \) of the original facial image \( X \), it returns a higher loss value; if it belongs to a misleading classification, it returns a lower loss value, thereby ensuring the privacy of the facial image \( X' \).

\[
L_{\text{adv}} = \mathbb{E}_{X'} \ell_f(X', t, t')
\]

(9)

In the white-box environment and with purposeless label training, the target face recognition network \( f \) can return a set of classification labels closest to the classification \( t \) of the original facial image \( X \) in Euclidean distance. One of the labels is selected for training to accelerate the GAN convergence rate.

Theoretically, the output range of the elements in the set \( U \) of left singular matrices is \([-1, 1]\). In actuality, the pixel value of each image varies considerably, and so the output range may be much smaller. As a result, as shown in (10), the perturbation range of the generator \( G \) output can be appropriately reduced:

\[
L_N = \mathbb{E}_U \left( \rho \|G(U)\|_2 \right)
\]

(10)

where \( \rho \) is the coefficient. Adjusting the size of \( \rho \) according to the value range of \( U \) can further reduce the GAN training time.

### B. NETWORK ARCHITECTURE

The structure of the privacy-protected facial image generation network based on the principal components of the adversarial segmented image blocks is shown in Figure 2. This is the so-called “PcadvGAN,” which includes a generator \( G \), a discriminator \( D \), and a target face adversarial recognition network \( f \). The left singular matrix set \( U \) of the facial image sub-block \( X_i \) was used as the input of the generator \( G \), as the GAN convergence rate is faster when the distribution of the initial data is similar to the distribution of the real data. The flow of the proposed method is shown in Table 1.

As shown in Table 1, the role of the image segmentation function \( \beta(x, num) \) is to segment the image matrix. The \( x \)
To ensure that the generated face privacy image $X'$ approximates the original facial image $X$ at the pixel level, as shown in (11), a pixel-level constraint is added, and the PSNR is utilized to evaluate the similarity of the two images.

$$L_{sim} = \max \left( \mathbb{E}_{X,X'} \left( \left( \frac{40 - PSNR(X,X')}{} \right) \right), 0 \right)$$  \hfill (11)

When the PSNR output value is higher than 40 dB, the two facial images are highly approximate [38], [39]. When the output value ranges from 30-40 dB, the generated image $X'$ is acceptable despite slight distortion [40]. To facilitate feedback from the loss function, the PSNR output can be restricted to $[-1, 1]$, that is, $L_{sim} \leq 0.25$ means that the generated facial image $X'$ is within an acceptable range.

The loss function of the entire PcadvGAN is:

$$L = L_{GAN} + \chi_1 L_{adv} + \chi_2 L_{N} + \chi_3 L_{sim}$$  \hfill (12)

where $\chi_1$, $\chi_2$, $\chi_3$ are the respective hyperparameters of the loss function.

V. EXPERIMENT

A. EXPERIMENTAL ENVIRONMENT

The software environment was a Windows10 64 bit operating system running the Google Tensorflow framework. Codes were written in Python. The VGGFACE2 and VGGFACE2 public datasets were used as the basis for the experiment. A black-box environment can be converted to a white-box environment by training distillation models and other methods [23], [24], so a white-box environment was comprised of an Intel i7-8700K CPU with 32GB DDR4 memory and an NVIDIA GeForce GTX 1080 graphics card.

The hardware environment for the experimental test was configured as follows: the operating system running the Google Tensorflow framework was a Windows10 64 bit system, and the NVIDIA GeForce GTX 1080 graphics card was used. The training set, development set, and test set were divided based on the 98:1:1 ratio. There are 2622 people with different identities in the VGGFACE datasets, and there are 2622 people with different identities in the VGGFACE datasets, and there are

B. NETWORK STRUCTURES OF GENERATOR AND DISCRIMINATOR

The discriminator uses three convolutional layers to extract the features of the input data. All layers use LeakyRelu as the activation function. Two of the convolutional layers use Instance Norm, thus improving the stability of the model and accelerating the convergence speed. The last fully-connected layer uses Sigmoid as the activation function. Figure 3 shows a diagram of this structure.

The generator network structure is shown in Figure 4. The input data passes through three convolutional layers and three deconvolutional layers and is output via the Tanh activation function. Instance Norm and LeakyRelu were imposed in each of the convolutional and deconvolutional layers here to enhance the robustness of the generator. LeakyRelu is an unsaturated activation function, which gives all negative values a non-zero slope to solve the problem of gradient disappearance. Four layers of residual blocks were also added between the convolutional layer and the deconvolutional layer to increase the generator’s network depth. The network structure of the residual block is shown in Figure 5. It consists of two convolutional layers and Instance Norm, and its activation function is Relu.

C. EXPERIMENT AND ANALYSIS OF FACIAL IMAGES

The VGGFACE and VGGFACE2 datasets are already properly trained in high-accuracy models [31]–[34]. The properly trained face recognition model was defined here as the target face recognition network $f$. The basic accuracy of each target face recognition network $f$ as-observed in this experiment listed in Table 2.

We divide the training set, development set, and test set according to the 98:1:1 ratio. There are 2622 people with different identities in the VGGFACE datasets, and there are
We take 2.55 million face images as the training set, 0.026 million face images as the development set, and 0.026 million face images as the test set. There are 9131 people with different identities in the VGGFACE2 datasets, and there are 3.31 million face images. We take 3.24 million face images as the training set, 0.033 million face images as the development set, and 0.033 million face images as the test set. The training process of PcadvGAN iterating 2,000 epochs is shown in Figure 6. L-noise represents the size of the perturbations $L_N$, and $L_{sim}$ is $L_{sim}$. The accuracy ($Acc$) is inversely proportional to the quality loss of facial image $L_N$. At the beginning of the training process, we set a large initial value for the perturbations. The values of $L_N$, $L_{sim}$, and $Acc$ vary widely. In order to reduce $L_N$ and $L_{sim}$, the accuracy rate of $Acc$ fluctuation frequency is very high. With the increase of training times, PcadvGAN gradually learned to balance the relationship between the three. After 600 training sessions, it is obvious that the fluctuation frequency of $Acc$, $L_N$ and $L_{sim}$ gradually decreases and finally reaches a stable state. Eventually, PcadvGAN learned to reduce the values of $L_N$ and $L_{sim}$ while maintaining an $Acc$ equal to 0. In figure 6, we highlighted two advantages in 2000 training sessions in green boxes.

After AdvGAN, PriGAN, and PcadvGAN iterated 2,000 epochs, the accuracy rate of the target face recognition network $f$ for generating face privacy images was as shown in Table 3. AdvGAN, PriGAN, and PcadvGAN both reduced the accuracy rate of the target face recognition network $f$ to 0%, which in practice would protect the privacy of the facial image.
In the case where the accuracy rate of the generated face privacy image Acc was reduced to 0% for the target face recognition network \( f \), the original facial image and the generated facial privacy images were taken to calculate the image similarity using (11) as shown in Figure 7.

Using Acc dropout as 0% as the benchmark, PriGAN \( L_{sim} > 0.25 \), the generated facial privacy images were visually unacceptable. When AdvGAN and PcadvGAN \( L_{sim} < 0.25 \), the generated facial privacy images were visually acceptable. Due to the loss function \( L_{sim} \) of PcadvGAN, the quality of these privacy images was better than those generated by AdvGAN and PriGAN.

The visual contrast between the generated facial privacy images and the original images are shown in Figure 8. Most features of the original facial images are retained and well visible to the human eye, thereby ensuring the “availability” of the generated image.

The visual contrast between facial images generated by AdvGAN, PriGAN, PcadvGAN, and original images are shown in Figure 9. The color and contrast ratio of facial images generated by PriGAN has changed a lot, and the chessboard effect of the images is obvious. The facial images generated by AdvGAN and PcadvGAN both can ensure the “availability.” Moreover, compared with AdvGAN and PriGAN, the quality of the facial images generated by PcadvGAN is better.

The pixel contrast between the generated and original facial images are shown in Figure 10. The perturbation interfered with the primary features of all faces and was not easily
filtered by the reverse denoising algorithm, which effectively protected the privacy of the generated facial images.

The training run times of AdvGAN, PriGAN, and PcadvGAN on 100 times batchsize=4 are shown in Figure 11. As the input matrix was segmented by the proposed method and compressed twice with PCA, the input parameters were actually 0.125 times those of the original image. PcadvGAN thus runs considerably faster than AdvGAN and PriGAN. However, as image segmentation and merging operations are lengthy and additional operations (e.g., batch size and pixel comparison) are required, the actual running speed was not as high as its theoretical speed.

D. ABBLATION STUDY

To confirm the contribution of the loss functions, we define two more loss functions for ablation study:

\[ L' = L_{GAN} + \chi_1 L_{adv} \]  \hspace{1cm} (13)
\[ L'' = L_{GAN} + \chi_1 L_{adv} + \chi_2 L_N \]  \hspace{1cm} (14)

The training process of PcadvGAN using (13) and (14) iterating 2,000 epochs is shown in Figure 12. L-noise represents the size of the perturbations \( L_N \), and L-sim is \( L_{sim} \).

When use loss functions (13), the values of \( L_N \) and \( L_{sim} \) stay in a high position without going down. This kind of training is totally ineffective. Due to the image proportion, the L-sim fluctuation is not obvious and actually fluctuates between 1.48 and 1.53. When use loss functions (14), \( L_{sim} \) reached the training bottleneck and remained stable around 0.3 in the middle and later period of training. The generated the facial privacy images can not guarantee practicability. Compared with Figure 12 and Figure 6, loss function (12) is the best way to generate the facial privacy images.

E. ROBUST

In order to test the robustness of the proposed method in this paper, PcadvGAN, AdvGAN, and PriGAN are used respectively, and the target face recognition network \( f \) without defense strategy is used to generate facial privacy images, and the generated results are respectively attacked against the target face recognition network \( f \) with different defense strategies, and the results of attack success rate are shown in Table 4.

It can be seen from Table 4 that when the defense strategy of the target face recognition network adopts the median...
The proposed method was designed to safeguard users’ privacy while ensuring the availability of their facial images. The method is superior to other similar methods in terms of image quality, running speed, and target recognition network accuracy.

ACKNOWLEDGMENT

The authors Jingjing Yang and Jiaxing Liu thank great help and encouragement from all members of the C508 Research Room, Hebei North University.

REFERENCES

[1] K. Guo, S. Wu, and Y. Xu, “Face recognition using both visible light image and near-infrared image and a deep network,” CAAJ Trans. Intell. Technol., vol. 2, no. 1, pp. 39–47, Mar. 2017.
[2] A. Sepas-Moghaddam, P. L. Correia, and F. Pereira, “Light field local binary patterns description for face recognition,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Beijing, China, Sep. 2017, pp. 3815–3819.
[3] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, “Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 1, pp. 156–171, Jan. 2017.
[4] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, and D. Yue, “Learning robust and discriminative low-rank representations for face recognition with occlusion,” Pattern Recognit., vol. 66, pp. 129–143, Jun. 2017.
[5] X. Fu, J. Lu, X. Zhang, X. Yang, and I. Unwala, “Intelligent in-vehicle safety and security monitoring system with face recognition,” in Proc. IEEE Int. Conf. Comput. Sci. Eng. (CSEE), IEEE Int. Conf. Embedded Ubiquitous Comput. (EUC), New York, NY, USA, Aug. 2019, pp. 225–229.
[6] R. Patil, S. Patil, S. Sagar, S. Sancheti, and S. More, “Secure online payment with facial recognition using CNN,” Int. Res. J. Eng. Technol., vol. 6, no. 4, pp. 604–607, 2019.
[7] S. Chen, S. Ding, H. Fu, Y. Xian, X. Liu, and C. Zhang, “Deep learning applied to smart home face recognition access control system,” in Proc. 2nd Int. Conf. Artif. Intell., Technol. Appl. (ICAITA), Chengdu, China, 2018, pp. 13–15.
[8] W. Shen, Z. Wu, and J. Zhang, “A face privacy protection algorithm based on block scrambling and deep learning,” in Proc. Int. Conf. Cloud Comput. Secure. Haikou, China: Springer, 2018, pp. 359–369.
[9] T. J. Schulz, “Schröms v. Data protection commissioner (CJEU),” Int. Legal Mater., vol. 56, no. 2, pp. 245–272, 2017.
[10] B. Henne and M. Smith, “Awareness about photos on the Web and how privacy-privacy-tradeoffs could help,” in Proc. Int. Conf. Financial Cryptogr. Data Secur. Okinawa, Japan: Springer, 2013, pp. 131–148.
[11] Z. Ma, Y. Liu, X. Liu, J. Ma, and K. Ren, “Lightweight privacy-preserving ensemble classification for face recognition,” IEEE Internet Things J., vol. 6, no. 3, pp. 5778–5790, Jun. 2019.
[12] Y. Wang and M. Kosinski, “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images,” J. Personality Social Psychol., vol. 114, no. 2, p. 246, 2018.
[13] X. Wu and X. Zhang, “Responses to critiques on machine learning of criminality perceptions,” Nov. 2016, arXiv:1611.04135. [Online]. Available: http://arxiv.org/abs/1611.04135
[14] W. Zeng and S. Lei, “Efficient frequency domain selective scrambling of digital video,” IEEE Trans. Multimedia, vol. 5, no. 1, pp. 118–129, Mar. 2003.
[15] J. S. Seo, S. O. Hwang, and Y.-H. Suh, “A reversible face de-identification method based on robust hashing,” in Int. Conf. Consum. Electron. Digit. Tech. Papers, Las Vegas, NV, USA, Jan. 2008, pp. 1–2.
[16] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent, “Large-scale privacy protection in Google street view,” in Proc. IEEE 12th Int. Conf. Comput. Vis., Kyoto, Japan, Sep. 2009, pp. 2–2173.
[17] L. Meng, Z. Sun, A. Arivaeenia, and K. L. Bennett, “Retaining expressions on de-identified faces,” in Proc. 37th Int. Conf. Inf. Commun. Technol., Electron. Microelectron. (MIPRO), Opatija, Croatia, May 2014, pp. 1252–1257.
[18] L. Meng and Z. Sun, “Face de-identification with perfect privacy protection,” in Proc. 37th Int. Conf. Inf. Commun. Technol., Electron. Microelectron. (MIPRO), Opatija, Croatia, May 2014, pp. 1234–1239.

VI. CONCLUSIONS

The paper proposes a method for protecting facial image privacy based on the principal components of the adversarial segmented image blocks. The proposed method was designed to safeguard users’ privacy while ensuring the availability of their facial images. The method is superior to other similar methods in terms of image quality, running speed, and target recognition network accuracy.
