A Fake Face Image Encryption With Style Mixing

Xin Yan, Shingo Otsu, and Taizo Suzuki
University of Tsukuba, Ibaraki, Japan
{yanxin,otsu}@wmp.cs.tsukuba.ac.jp, taizo@cs.tsukuba.ac.jp

Kosuke Shimizu
Gifu University, Gifu, Japan
shimizu.kosuke.x5@f.gifu-u.ac.jp

***Abstract*—In this study, we propose a novel perceptual encryption method for face images with style mixing via StyleGAN2 using the pixel2style2pixel (pSp) encoder. Our method generates natural fake face images by replacing specific latent vectors, ensuring privacy protection while allowing decryption by authorized users. Experimental results using CelebA-HQ dataset and Cross-Age Celebrity Dataset (CACD) demonstrate the effectiveness of our method in maintaining image quality and privacy.**

***Index Terms*—Fake face image, perceptual encryption, privacy protection, style mixing.**

I. Introduction

Image perceptual encryption [1] is a technology that allows encrypted information to be displayed as an image, unlike traditional encryption standards such as the Advanced Encryption Standard (AES) and Rivest–Shamir–Adleman (RSA). Commonly used encryption methods include sign inversion [2], which randomly inverts the signs of a signal, and shuffling [3], which randomly rearranges the signal at certain intervals. While most conventional methods adequately protect privacy, they have limitations. The encrypted image may appear "unnatural" due to obtrusive artifacts and colors, which can be unpleasant for observers and may attract the attention of malicious attackers.

With advances in image generation technology like generative adversarial networks (GANs) [4], it has become possible to generate highly realistic images. GANs are unsupervised learning models composed of two networks, the generator and the discriminator, which learn by competing against each other. StyleGAN2 [5], an extension of StyleGAN [6], is one of the most notable GANs. Especially for face images, it not only generates high-quality and high-resolution images but also allows for the separation of global attributes (facial outline, presence or absence of glasses, etc.) from local attributes (wrinkles, skin texture, etc.), which can be controlled. This separation enables style mixing, generating a new image by blending a content image with the style of another image (style image). Also, GAN inversions [7] find the latent code that most accurately reconstructs a given known image. While most methods provide better reconstruction quality than learned encoders, they often require significant computation time. The pixel2style2pixel (pSp) encoder [8] sacrifices less computation time and can be considered as one of the most practical inversion techniques.

In this study, we propose a "natural" fake face image encryption method with style mixing. Figure 1 shows a result of encryption and decryption with our method. Existing style mixing methods emphasize high-quality face replacement and do not consider encryption, which requires restoration (decryption) methods. Our method utilizes latent space via StyleGAN2 using the pSp encoder to achieve facial feature replacement and restoration.

Fig. 1. Resulting images with our method: (left to right) original, for authorized receiver, for third party, and for attacker.

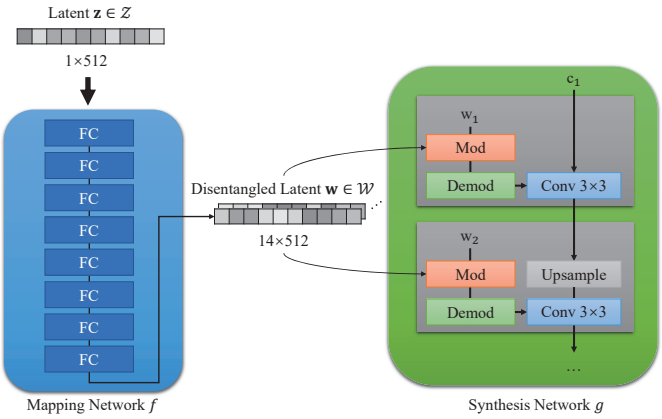

Fig. 2. Structure of StyleGAN2 (FC, Mod, Demod, and Conv mean fully connected layer, modulation layer, demodulation layer, and convolution layer, respectively).

II. Review

A. StyleGAN2

StyleGAN2 [5] as shown in Fig. 2 is a GAN improved the StyleGAN [6]. In conventional GAN, the input noise $\mathbf{z} \in \mathcal{Z}$ is directly input to the generator. However, when some combinations of image features do not exist, the latent space $\mathcal{Z}$ in which the input noise $\mathbf{z}$ exists may not be linear but a distorted space with entanglements. To address this problem, StyleGAN2 inputs $\mathbf{z}$ as a latent variable into a mapping network f consisting of 8-layer perceptrons, and creates an intermediate latent space $\mathcal{W}$. After mapping to $\mathcal{W}$ and obtaining another latent variable $\mathbf{w} \in \mathcal{W}$, the disentangled latent vector is input into the subsequent synthesis network g. Since f is learned to disentangle, it is possible to optimize $\mathbf{w}$ with less entangled image features. When training StyleGAN2,

979-8-3503-5506-2/24/$31.00 ©2024 IEEE

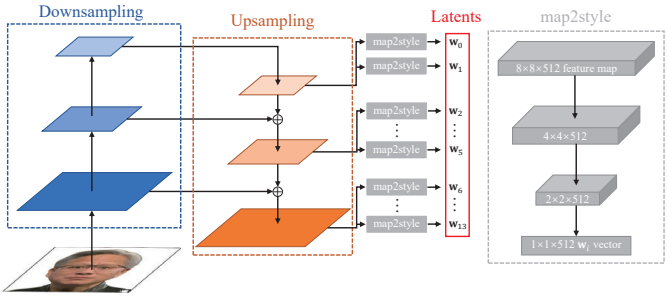

Fig. 3. Structure of pixel2Style2pixel encoder.

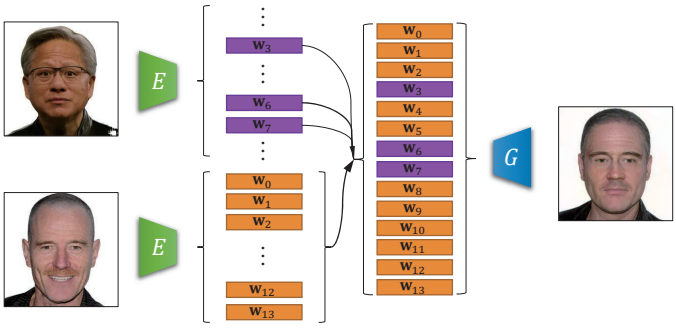

Fig. 4. Style mixing in the case of replacing the latent variables with indices except for $n = [3, 6, 7]$ (E and G mean encoder and generator, respectively).

a gradient penalty and a path length penalty are applied as loss functions.

B. pSp Encoder

The pSp encoder [8] as shown in Fig. 3 is one of the GAN inversions that enhance StyleGAN2 to generate more realistic images. It can accurately and efficiently embed facial features into the extended latent space $\mathcal{W}+$ without further optimization. First, it extracts feature maps using a standard feature pyramid on the ResNet [9] backbone. For each of the 18 target style, a small-scale mapping network is trained to extract the learned style from the corresponding feature map. Styles 0-2 are generated from small feature maps, styles 3-6 are generated from medium feature maps, and styles 7-17 are generated from the largest feature maps. The mapping network, termed as a map2style block, is a small-scale fully convolutional network that uses a series of two-stride convolutions followed by LeakyReLU [10] to gradually reduce the spatial size. Each of the generated 512 vectors is a latent variable and the resulting image is more similar to the real image. Additionally, the pSp encoder is defined as:

$$pSp(\mathbf{x}) = G(E(\mathbf{x}) + \overline{\mathbf{w}}) \tag{1}$$

where $\mathbf{x}$ is an input image, $\overline{\mathbf{w}}$ is an average style vector of a pre-trained generator, $E()$ is an encoder, and $G()$ is a generator. The encoder aims to learn latent codes in terms of average style vectors. When training the pSp encoder, the loss function is composed of pixel-wise $\mathcal{L}_2$ loss, LPIPS loss, regularization loss and cosine similarity.

C. Style Mixing

Style mixing generates a new image that reflects the style of a style image in a content image. Two different latent vectors are input into different levels of a synthesis network, so that the generated image contains features from both latent vectors. In StyleGAN2, two latent vectors $\mathbf{z}_1$ and $\mathbf{z}_2 \in \mathcal{Z}$ are sampled from the latent space, along with two intermediate latent vectors $\mathbf{w}_1$ and $\mathbf{w}_2 \in \mathcal{W}$. When inputting $\mathbf{w}_1$ and $\mathbf{w}_2$ as parameters for the regularization parts in the synthesis network, $\mathbf{w}_1$ is used up to a certain resolution scale, e.g., 4×4, and $\mathbf{w}_2$ is used for subsequent resolution scales. As a result, the synthesis network no longer learns that styles are correlated between layers at adjacent resolution scales, allowing the influence of styles to be localized to layers at each resolution scale. If the latent vectors of the content and style images are $\mathbf{w}^{\text{con}}$ and $\mathbf{w}^{\text{sty}}$, the style mixing for nth ($n = 0, 1, \cdots$) latent vector is defined as:

$$\mathbf{w}_n^{\text{con}} \leftarrow \mathbf{w}_n^{\text{con}} + \eta(\mathbf{w}_n^{\text{sty}} - \mathbf{w}_n^{\text{con}}), \tag{2}$$

where $\eta = [0, 1]_{\in \mathbb{R}}$ is the parameter that adjusts the degree of style application.

III. Fake Face Image Encryption With Style Mixing

A. Overview

We introduce a natural fake face image encryption with style mixing. When training StyleGAN2 and the pSp encoder, an additive angular margin loss (AAML) as described in [11] was newly added. If the latent vectors from the original and key face images are $\mathbf{w}^{\text{ori}}$ and $\mathbf{w}^{\text{key}}$, the style mixing for nth latent vector is defined as:

$$\mathbf{w}_n^{\text{ori}} \leftarrow \mathbf{w}_n^{\text{ori}} + (\mathbf{w}_n^{\text{key}} - \mathbf{w}_n^{\text{ori}}) = \mathbf{w}_n^{\text{key}}. \tag{3}$$

Here, note that η in (2) is set to 1 to enable the reconstruction. The flow of our method is outlined below (Fig. 5).

1) Sender Side:

i) Using the pSp encoder, a face image of the sender (original image) and an encryption key face image of the third party (encryption key face image) are converted into latent vectors.
ii) Using StyleGAN2, a natural encrypted image is generated by replacing some of the latent vectors of the original image with those of the encryption key face image.
iii) Using existing encoder, the encrypted image is sent to the receiver side.

2) Authorized Receiver Side:

i) Using existing decoder, the encrypted image is received from the sender side.
ii) Using the pSp encoder, the received encrypted image and a correct decryption key face image (another correct face image of the sender prepared in advance) are converted into latent vectors.
iii) Using StyleGAN2, the original image is reconstructed by replacing some of the latent vectors of the encrypted image with those of the decryption key face image.

3) Third Party and Attacker Sides: For a third party, the encrypted image is displayed as it is. For an attacker with an incorrect decryption key face image, different encrypted image is provided by replacing some of the latent vectors of the encrypted image with those of the decryption key face image.

IV. Experiments

A. Datasets and Training Settings

We used the CelebA-HQ dataset [12] and the Cross-Age Celebrity Dataset (CACD) [13] as image datasets. The CelebA-HQ dataset is a high-quality version of CelebA [14] and consists of $30,000$ images with 1024×1024 pixels. The CACD contains $163,446$ images from $2,000$ celebrities collected from the Internet. We randomly selected 90 % of the images from each dataset above as the training set and used the rest to evaluate the training results of the network. Note that as part of the preprocessing to simplify network training and experiments, we preprocessed all images as follows (Fig. 6): (i) image were resized from 1024×1024 to 256×256 pixels, (ii) the positions of eyes and mouth between images were aligned using InsightFace [15], and (iii) using a trimap that estimates foreground from background by inscribing unknown region, the foreground and background were separated (matted) using IndexNet [16].

B. Quantitative Evaluation Indicators

We used five quantitative evaluation indicators: peak signal-to-noise ratio (PSNR) [dB], structural similarity (SSIM) [17], learned perceptual image patch similarity (LPIPS) [18], and two additive angular margin losses (AAMLs) in ArcFace [11] and BlendFace [19] (denoted as AAML-A and AAML-B).

PSNR is an index that evaluates the mean pixel-by-pixel difference between input and output images and is defined as

$$\mathrm{PSNR} = 10 \log_{10} \frac{\mathrm{MAX}^2}{\mathrm{MSE}}, \tag{4}$$

where MAX is the maximum pixel value of the image and MSE is the mean square error of the input and output images.

SSIM is an index that indicates the structural similarity between the original image and the evaluation image, and is defined as

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{5}$$

where μ_x and μ_y are the averages in the local parts of the input and output images, σ_x and σ_y are the standard deviations in the local parts of the input and output images, σ_{xy} is the covariance of the input and output images, and C_1 and C_2 are the adjustment parameters.

LPIPS is an index that uses deep neural network, specifically AlexNet [20], to indicate similarity based on features of images being compared.

The AAML-A is a loss function introduced in ArcFace as follows:

$$\mathcal{L}_{\mathrm{Arc}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cdot(\cos(\theta_{y_i}+m))}}{e^{s\cdot(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{C} e^{s\cdot\cos(\theta_j)}}, \tag{6}$$

where N is the batch size, θ_{y_i} and θ_j are the angles between the feature vector of the i-th sample and the weight vector of their true class y_i and j, m is the angular margin, which is used to increase the inter-class distance, and s is the scale factor, which is used to control the magnitude of the logits. AAML-B, which is an extend version of AAML-A, has also been presented in BlendFace as follows:

$$\mathcal{L}_{\mathrm{Blend}} = \mathcal{L}_{\mathrm{adv}} + \mathcal{L}_{\mathrm{mask}} + \lambda_1\mathcal{L}_{\mathrm{Arc}} + \lambda_2\mathcal{L}_{\mathrm{rec}} + \lambda_3\mathcal{L}_{\mathrm{cyc}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{adv}}$ is the adversarial loss of GauGAN [21], $\mathcal{L}_{\mathrm{mask}}$ is the binary cross entropy loss, $\mathcal{L}_{\mathrm{rec}}$ is the reconstruction loss, and $\mathcal{L}_{\mathrm{cyc}}$ is the cyclic generation loss, which are defined as

$$\mathcal{L}_{\mathrm{adv}}(G, D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))] \tag{8}$$

$$\mathcal{L}_{\mathrm{mask}} = -\sum_{x,y}\Big\{M_{x,y}\log\widehat{M}_{x,y} + (1 - M_{x,y})\log(1 - \widehat{M}_{x,y})\Big\} \tag{9}$$

$$\mathcal{L}_{\mathrm{rec}} = \begin{cases} \|X_t - Y_{s,t}\|_1 & \text{if } ID(X_t) = ID(X_s), \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

$$\mathcal{L}_{\mathrm{cyc}} = \|X_t - G(X_t, Y_{s,t})\|_1, \tag{11}$$

respectively.

All indicators are calculated based on the original image.

C. Verification Results of Style Mixing

Note that at 256×256 pixels instead of 1024×1024 pixels, the latent space is 14×512 instead of 18×512, i.e., there are only 14 levels of latent space from $\mathbf{w}_0$ to $\mathbf{w}_{13}$. In order to determine the latent variables to be specified, Fig. 7 shows the verification results of an encrypted image in which only $\mathbf{w}_0$ to $\mathbf{w}_{13}$ latent variables were sequentially replaced by style mixing. For most patterns, even if $\mathbf{w}_8$ or higher was specified, the structural information did not change significantly and only the overall color tone changed. When $\mathbf{w}_4$ was specified, we can see that the facial expression changed to that of the encryption key face image and that the hairstyle also resembled that of the encryption key face image. We can also confirm that $\mathbf{w}_5$ inherited facial expression and hairstyle features from the encryption key face image, although not as much as $\mathbf{w}_4$. Based on these verifications, we summarized the transformation area according to the position of the latent variables in Table I. In this study, based on Table I, we replaced the latent variables with indices except for $n = [3, 6, 7]$.

D. Encryption and Decryption Evaluations

Table II shows quantitative evaluations of encryption and decryption. In addition, Fig. 8 and 9 shows the resulting images in encryption and decryption.

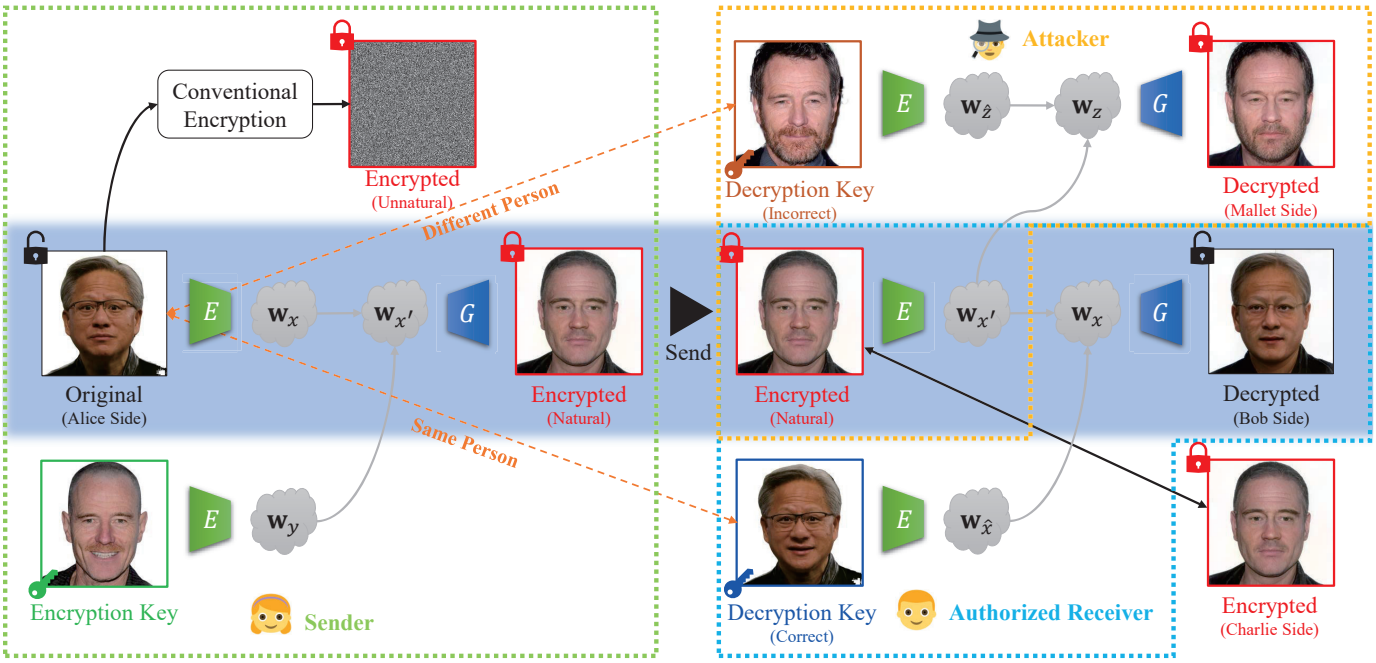

Fig. 5. Flow of our method with pSp encoder for E and StyleGAN2 for G (E, G, Alice, Bob, Charlie, and Mallet mean encoder, generator, sender, receiver, third party, and attacker, respectively).

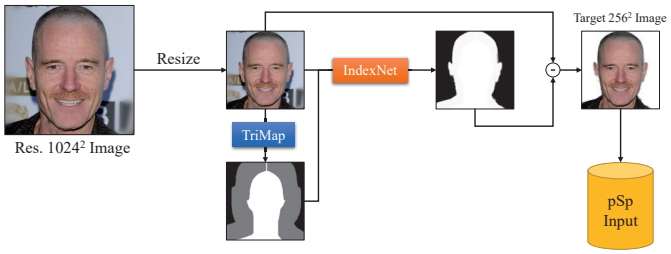

Fig. 6. Preprocessing for images.

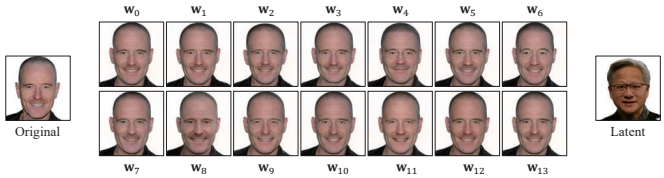

Fig. 7. Verification results of style mixing.

TABLE I
TARGET FOR REPLACING LATENT VARIABLES.

	$\mathbf{w}_0$	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	$\mathbf{w}_6$	$\mathbf{w}_7$	$\mathbf{w}_8$	$\mathbf{w}_{9+}$
Direction	✓	✓	✓	✓						
Shape				✓	✓	✓				
Expressions					✓	✓				
Hairstyle					✓	✓				
Eyes							✓	✓	✓	
Color etc.									✓	✓

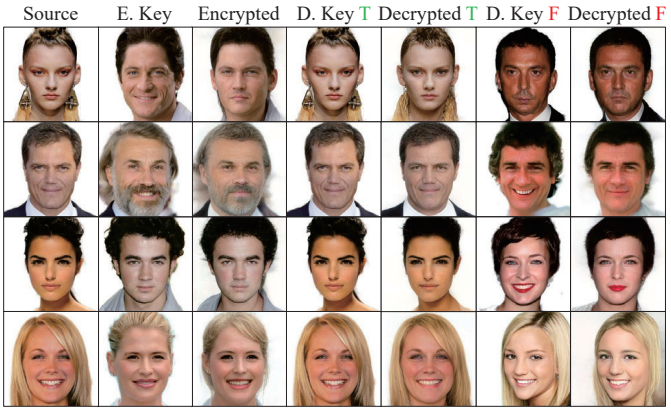

Fig. 8. Results of CelebA-HQ (Source, E. Key, Encrypted, D. Key T, Decrypted T, D. Key F, and Decrypted F mean original, encryption key, encrypted image, correct (true) decryption key, decrypted image with correct decryption key, incorrect decryption key, and decrypted image with incorrect (false) decryption key, respectively).

1) Encryption Evaluation: From Fig. 8 and Fig. 9, the encrypted images effectively retain the facial expressions of the original images while discarding all other features, thus protecting personal privacy. According to the results in Fig. II, the metrics based on pixel values and those based on discriminator remain at very low levels after image encryption, indicating that the encryption is highly effective.

2) Decryption Evaluation With Correct Decryption Key Face Image: For the CelebA-HQ dataset, whose the decryption key face images are same as the original images, the decrypted images were identical to the original images. While this is an impossible situation in real life, we evaluated an experiment to verify the effectiveness of decryption as the first step.

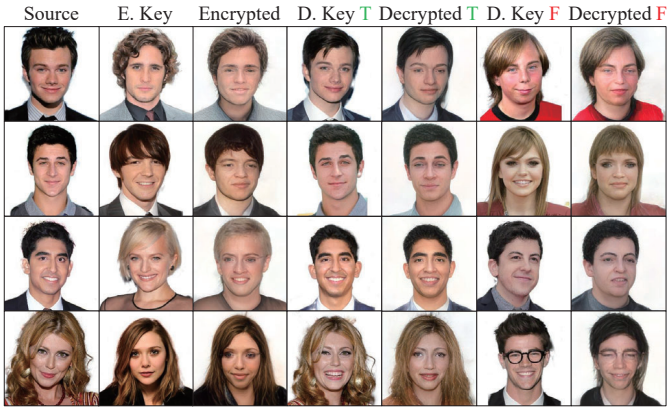

Fig. 9. Results of CACD (Source, E. Key, Encrypted, D. Key T, Decrypted T, D. Key F, and Decrypted F mean original, encryption key, encrypted image, correct (true) decryption key, decrypted image with correct decryption key, incorrect decryption key, and decrypted image with incorrect (false) decryption key, respectively).

TABLE II
QUANTITATIVE EVALUATION: (A) APPROXIMATED IMAGE, (B) ENCRYPTED IMAGE, (C) DECRYPTED IMAGE WITH CORRECT DECRYPTION KEY FACE IMAGE, AND (D) DECRYPTED IMAGE WITH INCORRECT DECRYPTION KEY FACE IMAGE.

	CelebA-HQ				CACD			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
PSNR↑	16.07	8.99	16.07	10.00	15.59	7.50	15.59	7.2
SSIM↑	0.774	0.600	0.774	0.599	0.752	0.561	0.752	0.597
LPIPS↓	0.307	0.547	0.307	0.488	0.310	0.662	0.310	0.666
ArcFace↑	68.12	6.36	68.12	7.75	68.22	5.39	68.22	2.24
BlendFace↑	56.49	8.20	56.49	1.37	59.60	7.02	59.60	1.47

The quality of the decrypted image was slightly degraded compared to the original image, but the higher the reproducibility of the approximate image, the higher the quantitative evaluation value recorded, indicating a certain level of decryption effect. However, the generated results were not exactly the same as original image, but from the qualitative analysis, there was not much difference between the two. The cause is that although the original image and the plain text key face image are the same person, each part such as the eyes and mouth changes slightly depending on the facial expression.

3) Decryption Evaluation With Incorrect Decryption Key Face Image: According to Fig. 5, when a third party attempts to decrypt the received image using a incorrect decryption key face image, a natural face image with features from the incorrect decryption key face image that is different from the original image will be generated. From Table II, the difference between the encrypted image and the image decrypted using the incorrect decryption key is not obvious. In other words, our experiment verified the security of the encryption process and ensured that a third party cannot decrypt the image even if it tries to reconstruct, without the correct decryption key.

V. CONCLUSION

In this study, we proposed a natural fake face image encryption based on style mixing. We selected StyleGAN2 with the pSp encoder and discussed a novel fake face image perceptual encryption that is different from conventional methods and the restoration (decryption) of the original image. Through encryption experiments using our method, we demonstrated its effectiveness and identified some issues.

REFERENCES

[1] S. Li, *Perceptual Encryption of Digital Images and Videos*, CRC Press, 2013.
[2] B. Zeng, S.-J. A. Yeung, S. Zhu, and M. Gabbouj, "Perceptual encryption of H. 264 videos: Embedding sign-flips into the integer-based transforms," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 2, pp. 309–320, Feb. 2014.
[3] R. Durstenfeld, "Algorithm 235: Random permutation," *Communications of the ACM*, vol. 7, no. 7, pp. 420, 1964.
[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montreal, Canada, 2014, pp. 672–2680.
[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE CVPR*, Virtual, Jun. 2020, pp. 8110–8119.
[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 4401–4410.
[7] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "GAN inversion: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3121–3138, Mar. 2023.
[8] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A StyleGAN encoder for image-to-image translation," in *Proc. IEEE CVPR*, Virtual, Jun. 2021, pp. 2287–2296.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, Jun.–Jul. 2016, pp. 770–778.
[10] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. IEEE ICML*, Atlanta, GA, Jun. 2013, vol. 30, p. 3.
[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 4690–4699.
[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. IEEE ICLR*, Vancouver, Canada, Apr.–May 2018, pp. 4401–4410.
[13] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. ECCV*, Zurich, Switzerland, Sep. 2014, pp. 768–783.
[14] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE ICCV*, Santiago, Chile, Dec. 2015.
[15] J. Guo, J. Deng, N. Xue, and S. Zafeiriou, "Stacked dense U-Nets with dual transformers for robust face alignment," in *Proc. BMVC*, Newcastle, UK, Sep. 2018.
[16] H. Lu, Y. Dai, C. Shen, and S. Xu, "Indices matter: Learning to index for deep image matting," in *Proc. IEEE ICCV*, Seoul, Korea, Oct.–Nov. 2019.
[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 586–595.
[19] K. Shiohara, X. Yang, and T. Taketomi, "BlendFace: Re-designing identity encoders for face-swapping," in *Proc. IEEE ICCV*, Paris, France, Oct. 2023, pp. 7634–7644.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, Lake Tahoe, NV, USA, Dec. 2012, vol. 25.
[21] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 2337–2346.