

# Cancelable Face Recognition Using Deep Steganography

Koichi Ito, *Member, IEEE*, Takashi Kozu, Hiroya Kawai, Goki Hanawa, and Takafumi Aoki, *Senior Member, IEEE*

**Abstract**—In biometrics, the secure transfer and storage of biometric samples are important for protecting the privacy and security of the data subject. One of the methods for authentication while protecting biometric samples is cancelable biometrics, which performs transformation of features and uses the transformed features for authentication. Among the methods of cancelable biometrics, steganography-based approaches have been proposed, in which secret information is embedded in another to hide its existence. In this paper, we propose cancelable biometrics based on deep steganography for face recognition. We embed a face image or its face features into a cover image to generate a stego image with the same appearance as the cover image. By using a dedicated face feature extractor, we can perform face recognition without restoring the embedded face image or face features from the stego image. We demonstrate the effectiveness of the proposed method compared to conventional steganography-based methods through performance and security evaluation experiments using public face image datasets. In addition, we present one of the potential applications of the proposed method to improve the security of face recognition by using a QR code with a one-time password for the cover image.

**Index Terms**—cancelable biometrics, face recognition, steganography, biometrics, template protection

## I. INTRODUCTION

**B**IOMETRIC recognition is a technology for identifying individuals using their physical or behavioral characteristics, and is used for user authentication in smartphones and identity verification in immigration control due to its high accuracy and convenience [1]. Face recognition is more convenient and cost effective than other biometric recognition such as fingerprint recognition and iris recognition since face recognition does not require special equipment to acquire face images [1], [2]. In addition, with the recent significant development of deep learning, face recognition performance has been dramatically improved [3]–[5]. Although many researches focusing on recognition performance have been reported, the recent challenge is to improve the security for biometric recognition systems to prevent attacks from malicious third parties. In particular, the protection of biometric samples has become a crucial issue since they are irreplaceable and unique personal information. The same biometric samples must be registered in different systems, resulting in privacy and security vulnerabilities. The biometric samples may be compromised by attacks on the system through the network, resulting in a loss of biometric authentication of a data subject using them. It is necessary to protect the biometric

samples in biometric recognition systems because of the risk of presentation attacks and privacy violations using the leaked biometric samples. In this paper, we focus on security issues in face recognition.

One of the approaches to protect biometric samples is cancelable biometrics [6], [7], also known as revocable biometrics [1], which converts biometric samples into a different format and uses the converted features in biometric recognition while protecting the original biometric samples. The use of cancelable biometrics allows individuals to register and revoke different biometric templates. The major methods of cancelable biometrics generate a cancelable template by distorting the template through image deformation and feature encryption [7]. The security concerns of cancelable biometrics are that the distortion parameters need to be safely managed for each template, and that the image is known to be protected by the distortion of image and other factors.

A new approach of cancelable biometrics is considered using steganography [8]. Steganography is one of the techniques to keep the information secret by embedding it in other information [9]. Steganography is especially used in the field of confidential communications since it can be used to transfer confidential information only to those who know how to extract the embedded information. Steganography cannot be applied directly to cancelable biometrics since the embedded secret information is extracted as it is. Choudhury et al. [8] encode the iris image to be hidden by Huffman coding, embed it in the DCT coefficients of another iris image, and generate a protected template by inverse DCT of them. They did not discuss the embedding of secret information for arbitrary images and verification with protected templates. Another problem with steganography is that the amount of embedded information is limited. To address this problem, Deep Steganography (DS) using Convolutional Neural Network (CNN) has been proposed, which can embed the information of a single image into another image of the same size using CNN [10]. Using this characteristic, Multitask Identity-Aware Image Steganography (MIAIS) has been proposed to embed a face image into an arbitrary image and extract features corresponding to the face image [11]. MIAIS is not secure as a face recognition method since the quality of the embedded image is low and the face image is visible in the embedded image.

In this paper, we propose a novel method for cancelable face recognition using DS [10]. The proposed method consists of Hiding Network (HN) and Extracting Network (EN). In HN, the secret information, e.g., a face image and face features, and the cover image for embedding the secret information are

K. Ito, T. Kozu, H. Kawai, G. Hanawa, and T. Aoki are with Graduate School of Information Sciences, Tohoku University, 6-6-05, Aramaki Aza Aoba, Aoba-ku, Sendai-shi, Miyagi, 9808579, Japan (e-mail: ito@aoki.ecei.tohoku.ac.jp).

input, and the stego image in which the secret information is embedded is output, where the appearance of the stego image is the same as the cover image. In EN, face features to be verified are extracted from the stego image without extracting the embedded secret information unlike DS. We introduce three loss functions: the reconstruction loss, perceptual loss, and feature reconstruction loss, to train the proposed network. The reconstruction loss is defined as the difference per pixel between the cover and stego images. The perceptual loss is defined as the difference between high-level image feature representations extracted from pre-trained CNNs [12]. The feature reconstruction loss is defined as the difference between face features extracted by a face feature extractor and EN. Unlike existing cancelable biometrics, which require the users to manage user-specific parameters, the proposed method requires managing EN at the server, and does not require the users to manage any parameters. The proposed method can generate higher-quality stego images than MIAIS without sacrificing the face recognition accuracy with the introduction of the three loss functions. We demonstrate the effectiveness of the proposed method compared to DS [10] and MIAIS [11] through performance and security evaluation experiments using the Labeled Faces in the Wild (LFW) dataset<sup>1</sup>, the CASIA-WebFace (CASIA) dataset [13], and the Asian Face Dataset<sup>2</sup> (AFD) [14], which are large-scale public datasets of face images. In addition, one of the potential applications of the proposed method is presented. The QR code image generated from the one-time password (OTP) can be used as the cover image of the proposed method to perform two-factor authentication, which can improve the security without sacrificing the usability.

The contributions of this paper are summarized below:

- propose a novel method for face recognition with cancelable biometrics using steganography,
- design a novel network for embedding face images and extracting face features, inspired by DS, and
- achieve that the recognition accuracy of the proposed method is comparable to that of face recognition methods without cancelable biometrics.

## II. RELATED WORK

In this section, we briefly summarize the related work on cancelable biometrics, steganography, and face recognition, which are the fundamental techniques used in the proposed method.

### A. Cancelable Biometrics

There is a risk of identity theft and privacy violation once the biometric template is leaked since it cannot be changed, unlike passwords. Cancelable biometrics converts biometric samples into a different format and uses the converted features in biometric recognition while protecting the original biometric samples [6], [7]. Since the original biometric samples cannot be recovered even if the transformed features are leaked,

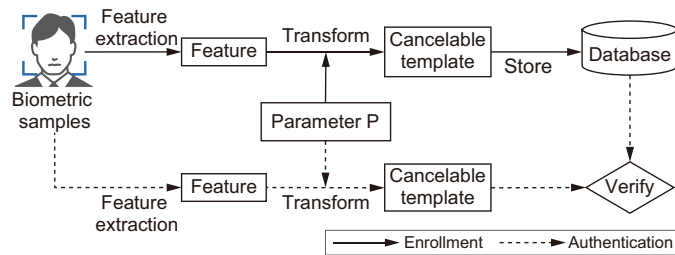


Fig. 1. Processing flow of typical methods of cancelable biometrics.

the use of cancelable biometrics allows us to protect the privacy of the data subject. Cancelable biometric algorithms need to satisfy biometric template protection criteria: (i) diversity, (ii) unlinkability, (iii) revocability/reusability, (iv) non-invertibility/irreversibility, and (v) performance preservation as stated in [7], [15], [16] and ISO/IEC 24745 standard<sup>3</sup>:

- Diversity: Multiple different templates can be created from the same biometric samples,
- Unlinkability: It should be difficult to link multiple templates of the same data subject across different applications or databases,
- Revocability/Reusability: If a cancelable template is compromised, it should be possible to revoke it and issue a new one that is not linked to the old one,
- Non-invertibility/irreversibility: Original biometric samples cannot be recovered if the generated template got compromised, and
- Performance preservation: The recognition performance of the cancelable biometric system should be comparable to that of a system using the original biometric data.

So far, there have been several methods proposed to protect biometric samples and most methods follow the processing flow as shown in Fig. 1. A cancelable template is generated using some transformation parameters for the features extracted from the biometric sample, and registered in the database. During authentication, a cancelable template is generated for an input in the same way, and the person is identified by matching it with the cancelable templates registered in the database. There are several methods proposed to protect biometric samples by deforming and encrypting images. For example, morphing is used to transform biometric samples by curving face images [17], and visual cryptography is used to make biometric samples imperceptible [18]. Deforming the image or encrypting the features makes it known that the cancelable template includes secret information, and thus may motivate a third party to attack. Other methods have been proposed to create cancelable templates by an inner product of user-specific random numbers seeded with PINs and biometric data [19], or by convolution of user-specific random numbers and biometric images [20]. When a data subject needs to manage auxiliary information such as PIN or secret key in addition to biometric samples, there is a risk of forgetting, losing, or stealing them as in the case of conventional password-based authentication. Steganography

<sup>1</sup><http://vis-www.cs.umass.edu/lfw/>

<sup>2</sup><https://github.com/X-zhangyang/Asian-Face-Image-Dataset-AFD-dataset>

<sup>3</sup><https://www.iso.org/standard/75302.html>

is one of the methods to protect biometric samples without being suspected by a third party. Cancelable biometrics using steganography for iris recognition has been proposed [8], while they did not discuss the embedding of secret information for arbitrary images and verification with protected templates.

### B. Steganography

Steganography is a technique to keep the information secret by embedding it in other information and can be classified into two major approaches. The first approach is to embed the information in the spatial domain of the image [21]–[24]. For example, there is a method to embed the information in locations where the human cannot detect bit reversals, such as the lower 1 bit or 2 bits of each pixel [24]. Another approach is to embed the information in the frequency domain of the image [25], [26]. Embedding methods in the frequency domain are more robust to image compression and resizing than methods in the spatial domain, however, they have the disadvantage of embedding a small amount of information. Recently, CNN-based methods have been proposed for steganography [10], [27]–[29], which can embed a greater amount of information than steganography without CNN. One of the most popular methods is DS [10], which can embed another image of the same size directly into one image as shown in Fig. 2. The stego image is generated by embedding the image to be hidden into another image (cover image) in Hiding Network (HN), where the stego image is indistinguishable from the cover image. The hidden image can be extracted by inputting the stego image into Revealing Network (RN). In our proposed method, stego images are generated inspired by the DS framework.

### C. Face Recognition

Face recognition is more convenient and cost effective than other biometric recognition since face recognition does not require special equipment to acquire face images [1]. With the development of deep learning, the accuracy of face recognition has improved significantly by CNN enough to make it practical [3]–[5]. Taigman et al. [3] proposed a face recognition method called DeepFace and demonstrated that face features extracted by CNN trained on 4.4 million labeled face images consisting of 4,030 individuals can provide recognition accuracy comparable to that of humans. Recently, the accuracy of face recognition has been improved by using deep metric learning, which trains CNNs so that the within-class variance is small and the between-class variance is large in the feature space. Schroff et al. [4] proposed FaceNet using triplet loss [30]. Triplet loss is a loss function that trains a reference feature to be closer in the Euclidean distance to features of the same class and farther in the Euclidean distance to features of different classes. Deng et al. [5] proposed a face recognition method using ArcFace. ArcFace is a loss function that trains a feature to have a higher cosine similarity with features of the same class and a lower cosine similarity with features of different classes by adding a margin only to the positive class. Multitask Identity-Aware Image Steganography (MIAIS), which is a combination of steganography and face recognition, has been proposed [11]. MIAIS consists of a steg-generator to hide the

secret images into the cover images, two steg-classifiers to recognize the container images, i.e., the stego images in this paper, and a steg-restorer to restore the secret images from the container images. This network is trained using the min-max optimization strategy for the five loss functions: visual similarity loss, discrepancy loss, recognition loss, content loss, and restoration loss functions. MIAIS is not necessarily a secure face recognition method since the quality of the stego image is low and the face image is visible in the stego image.

## III. CANCELABLE FACE RECOGNITION USING DEEP STEGANOGRAPHY

We describe a face recognition method with cancelable biometrics using DS proposed in this paper. We can simply embed the face image to be recognized into an arbitrary cover image using DS. In this case, the face image is embedded in HN, and the embedded face image is extracted by RN, and then the extracted image is inputted into the feature extractor for face recognition. This simple method violates one of the requirements of cancelable biometrics, i.e., non-invertibility, since the face image is extracted from the stego image during the processing of face recognition. Therefore, we propose a face recognition method that satisfies the requirements of cancelable biometrics by extracting face features from the stego image without extracting any information embedded in HN.

Fig. 3 shows an overview of the proposed method consisting of Hiding Network (HN) and Extracting Network (EN). In both the enrollment and authentication phases, a face image (or a face feature vector)  $F$  and a cover image  $C$  are concatenated in the channel direction and input to HN to generate a stego image  $S$  whose visual appearance is the same as the cover image  $C$ . Note that the cover image  $C$  can be selected in the different images for both phases. In the enrollment phase, the generated stego images  $S_r$  are stored in the database. In the authentication phase, the stego image  $S_r$  taken from the database and the stego image  $S_i$  generated from  $F_i$  are input to EN, respectively, and the feature vectors  $f_r$  and  $f_i$  are extracted. As described later, the feature vectors  $f_i$  and  $f_r$  extracted by EN are different from those of  $F_r$  and  $F_i$  input to HN, since the feature vectors  $f_r$  and  $f_i$  is mixed using a pseudo-random matrix in training of EN. The matching is then based on the similarity between the feature vectors  $f_r$  and  $f_i$ . In this method, the data subject does not need to keep any user-specific transformation parameters. Since the stego image, which is a template, has the same appearance as an arbitrary cover image, it is not known to be secret information. In addition, it is difficult to recover the original face image or face features from the stego image, since the transformed features, not the face image, are retrieved from the stego image.

Fig. 4 illustrates the details of the network architecture of HN used in the proposed method. The network architecture of HN is based on U-Net [31], which is widely used as an encoder-decoder network in the field of image processing. U-Net consists of an encoder and a decoder, and skip connections are introduced between the encoder and the decoder. The skip connections propagate features from the encoder to the

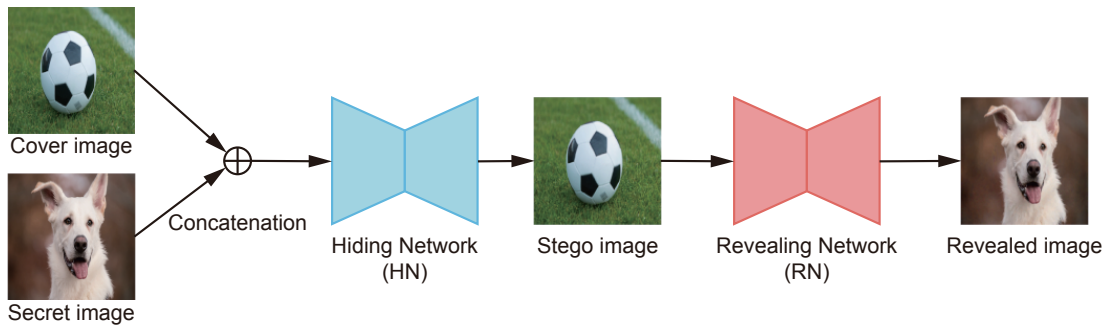


Fig. 2. Overview of Deep Steganography (DS) [10], which consists of Hiding Network (HN) and Revealing Network (RN). HN generates the stego image in which the secret image is embedded in the cover image. RN restores the secret image from the stego image.

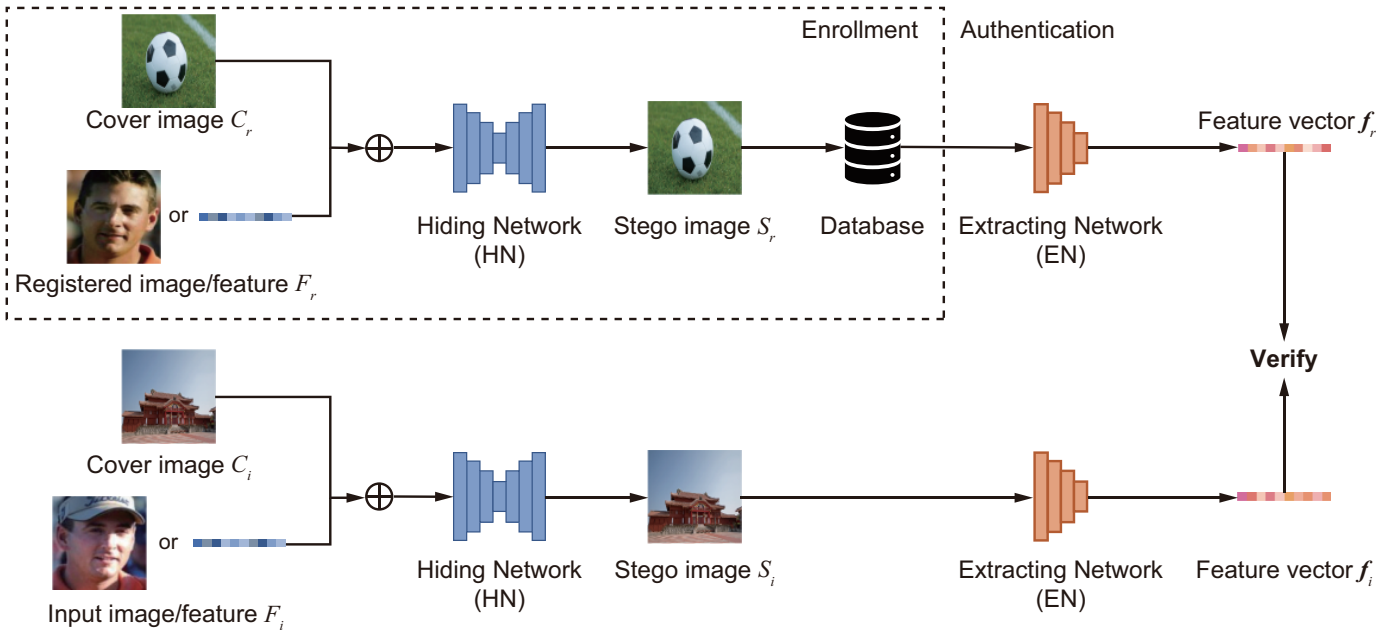


Fig. 3. Overview of face recognition with cancelable biometrics using deep steganography proposed in this paper. The proposed method consists of Hiding Network (HN) and Extracting Network (EN). HN generates the stego image in which the secret information, e.g., face image or face features, is embedded in the cover image. EN extracts face features from the stego image without restoring the secret information.

decoder, thus producing a high-resolution image and preventing gradient vanishing. HN of the proposed method employs residual blocks of ResNet [32] as encoder, which is different from HN of DS. In residual blocks, the skip connection is also used to avoid gradient vanishing. The input to HN is a 6-channel tensor that concatenates the face image (or face features) and the cover image in the channel direction. In the case of inputting face features and a cover image, the 512-dimensional face features are duplicated up to the total number of dimensions of the cover image, i.e.,  $256 \times 256 \times 3$ , and then reshaped to obtain an image with  $256 \times 256$  pixels and 3 color channels, which is the same as the cover image.

Unlike RN in DS, EN extracts features for face recognition from stego images. Fig. 5 illustrates the details of the network architecture of EN used in the proposed method. The network architecture of EN consists of IRBlocks<sup>4</sup>, which are an improved version of ResNet-18 [32] used in ArcFace. In EN,

the skip connections are also used to avoid gradient vanishing. Squeeze-and-Excitation block (SE block) [33] is added at the end of each block to propagate important information for CNN training. SE block extracts the representative value of each channel in the adaptive average pooling layer, obtains the weights of each channel by passing through the fully-connected layer and the activation layer, and multiplies them with the input to propagate the important features.

Three loss functions are used to train HN and EN in the proposed method as shown in Fig. 6. The first loss function is the reconstruction loss  $L_{rec}$ , which makes the stego image  $S$  equal to the cover image  $C$ , and is defined by

$$L_{rec} = \frac{1}{N} \sum_{k=1}^N \|C_k - S_k\|_2^2, \quad (1)$$

where  $k$  indicates the image number, and  $N$  is the number of images per batch. The stego image  $S$  is obtained by inputting the 6-channel data, which is the concatenation of the face image (or face features)  $F$  and the cover image  $C$  as described

<sup>4</sup><https://github.com/ronghuaiyang/arcface-pytorch/blob/master/models/resnet.py>

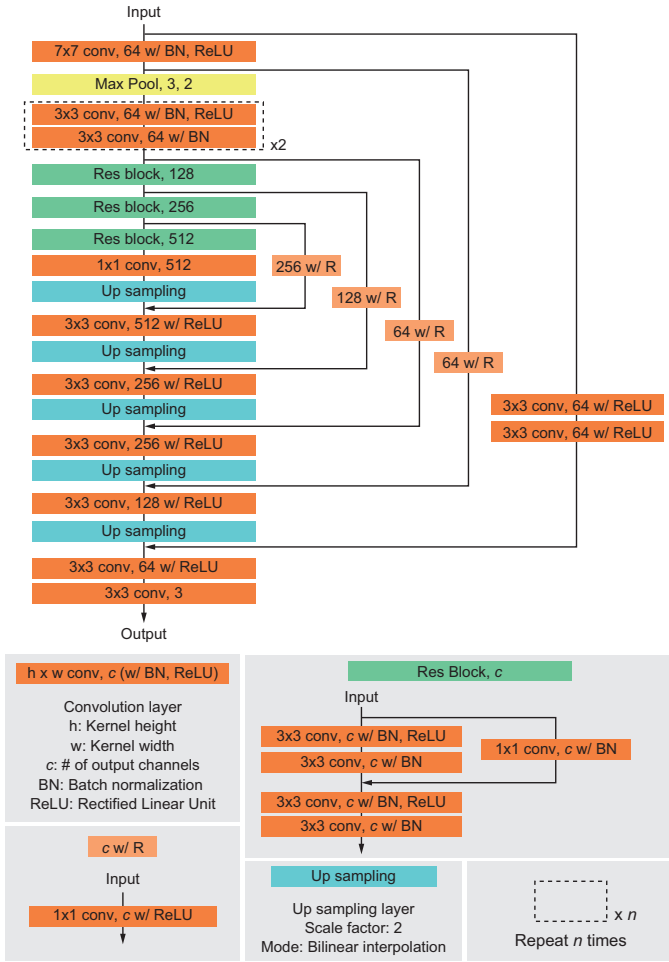


Fig. 4. Network architecture of Hiding Network (HN) used in the proposed method.

above, into HN. The second loss function is the perception loss  $L_{perc}$ , which makes the image features of the cover image  $C$  and the stego image  $S$  closer together, where we employ the trained VGG-19 [34] as a feature extractor [12], and is defined by

$$L_{perc} = \frac{1}{N} \sum_{k=1}^N \|\mathbf{f}_{vggk} - \mathbf{f}'_{vggk}\|_1, \quad (2)$$

where  $\mathbf{f}_{vgg}$  and  $\mathbf{f}'_{vgg}$  are image features extracted by VGG-19 pretrained on ImageNet. We introduce  $L_{perc}$  into the proposed method to improve the quality of stego images, which is one of the issues in DS and MIAIS. The third loss function is the feature loss  $L_{feat}$ , which makes the feature  $\mathbf{f}'_{face}$  obtained by inputting the stego image  $S$  to EN equal to the feature  $\mathbf{f}_{face}$  obtained by inputting the face image  $F$  into the face feature extractor, and is defined by

$$L_{feat} = \begin{cases} \frac{1}{N} \sum_{k=1}^N \|\mathbf{f}_{facek} - \mathbf{f}'_{facek}\|_2^2 & \text{(FaceNet)} \\ \frac{1}{N} \sum_{k=1}^N \{1 - \cos(\mathbf{f}_{facek}, \mathbf{f}'_{facek})\} & \text{(ArcFace)} \end{cases}, \quad (3)$$

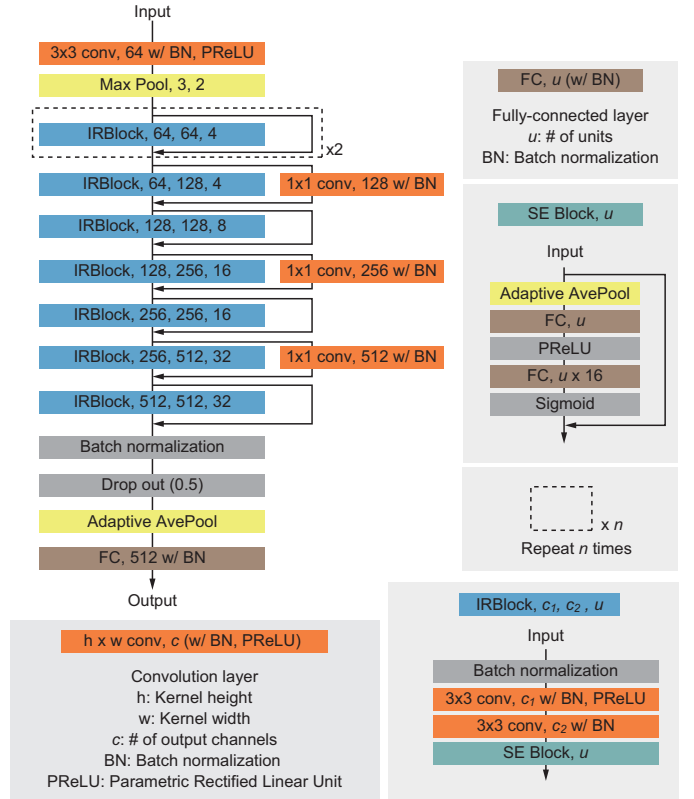


Fig. 5. Network architecture of Extracting Network (EN) used in the proposed method.

where  $\cos(\mathbf{f}_{facek}, \mathbf{f}'_{facek})$  indicates the cosine similarity between feature vectors  $\mathbf{f}_{facek}$  and  $\mathbf{f}'_{facek}$ . Note that  $L_{feat}$  depends on the type of face feature extractors used for training as in Eq. (3). In this paper, CNNs for face recognition trained with triplet loss [4] and ArcFace [5] are called FaceNet and ArcFace, respectively. As shown in Fig. 6 (c),  $L_{feat}$  mixes the features obtained from the face feature extractor by multiplying them by a pseudo-random matrix. We train CNN using a fixed pseudo-random matrix, and then discard the pseudo-random matrix when training is finished. This makes it difficult to recover the original face image or face features from the features output from EN, and therefore satisfies non-invertibility of cancelable biometrics. It can generate a variety of features by changing the pseudo-random matrix each training, and therefore satisfies diversity and revocability of cancelable biometrics. The overall loss function  $L$  is defined as the sum of the three loss functions and is given by

$$L = L_{rec} + \alpha \cdot L_{feat} + \beta \cdot L_{perc}, \quad (4)$$

where  $\alpha$  and  $\beta$  indicate hyperparameters.

#### IV. EXPERIMENTS AND DISCUSSION

We describe experiments to evaluate the performance of the proposed method in face recognition and the security of stego images.

##### A. Datasets

We describe the four public face image datasets used in the experiments.

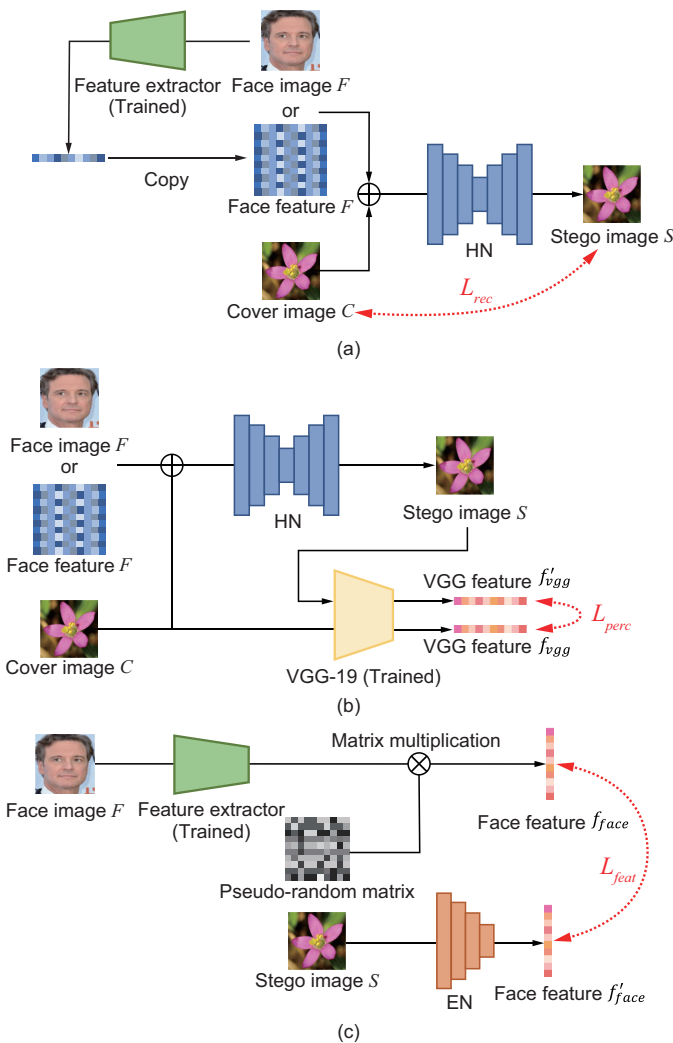


Fig. 6. Overview of loss functions in the training process: (a) reconstruction loss, (b) perception loss, and (c) feature loss.

### (i) CelebFaces Attributes Dataset (CelebA) [35]

CelebA consists of 202,599 face images taken from 10,177 people collected on the Internet. Some of the images in this dataset contain images that do not face the front or are partially occluded. In the experiments, CelebA is used to train CNNs. We use 199,599 images for training and the remaining 3,000 images for validation.

### (ii) Labeled Faces in the Wild (LFW) [36]

LFW consists of 13,233 face images taken from 5,749 people collected on the Internet. As in CelebA, some of the images in LFW contain images that do not face the front or are partially occluded. In the experiments, LFW is used to evaluate the performance of face recognition. We follow the evaluation protocol recommended by LFW, which uses 3,000 pairs each of genuine and impostor pairs.

### (iii) CASIA-WebFace Dataset (CASIA) [37]

CASIA consists of 494,414 face images taken from 10,575 people collected on the Internet. We use 490,740 of these images whose faces can be detected by MTCNN [38]. As with other datasets, some of the images do not face the front, and some of the faces are occluded. In our experiments, CASIA

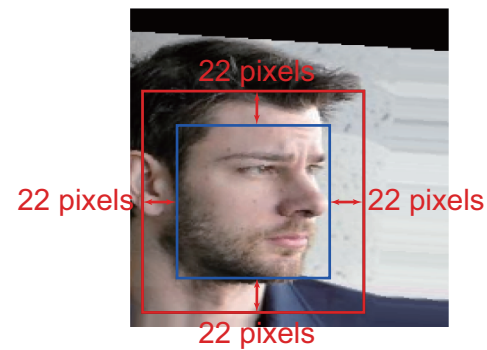


Fig. 7. Face region detected by MTCNN (blue rectangle) and enlarged face region used in the experiments (red rectangle).

is used to evaluate the performance of face recognition. For each person, the 21,150 pairs created by randomly selecting two pairs of the same person are used as genuine pairs, and the 21,150 pairs created by randomly selecting two different persons are used as impostor pairs.

### (iv) Asian Face Dataset (AFD) [14]

AFD consists of 360,000 face images taken from 2,019 people collected on the Internet. This dataset consists of Asian people and contains images of faces that do not face the front or are partially occluded. In this experiment, we use the training and test datasets provided on the website<sup>5</sup> for training CNNs and evaluating the performance of face recognition, respectively. The training dataset contains 1,662 face images, and the test dataset contains 357 face images. For the performance evaluation, we use a pair of face images selected from a randomly chosen person as the genuine pair, and at the same time, a different person is selected as the impostor pair. The above procedure is repeated 2,000 times to create 2,000 genuine pairs and 2,000 impostor pairs, respectively.

Face detection is performed on the face images in each dataset using MTCNN [38], and only face regions are extracted. As shown in Fig. 7, the face detection area is expanded by 22 pixels around the area to include the forehead, hair, and ears. Among the face image datasets, there were some images whose faces could not be detected by MTCNN only in CASIA, therefore, those face images were excluded from the experiments. The images are then resized to  $256 \times 256$  pixels and used as input for CNNs. Fig. 8 shows an example of the face images included in each dataset.

## B. Experiment Setup

The following is a summary of the experiment setup of this paper. HN and EN of the proposed method are trained using CelebA, unless otherwise mentioned. As mentioned above, 199,599 images are used for training and the remaining 3,000 for validation. The 6-channel data, which concatenates the face image  $F$  and the cover image  $C$  in the channel direction, is input to HN. In the case of using face features instead of face images, face features are extracted from the face image using a face feature extractor as described above, duplicated until they have the same size as the cover image, and concatenated with

<sup>5</sup><https://github.com/X-zhangyang/Asian-Face-Image-Dataset-AFD-dataset>



Fig. 8. Examples of face images used in the experiments: (a) CelebA, (b) LFW, (c) CASIA, and (d) AFD.

the cover image in the same way. This is why the proposed method embeds secret information of the same size as the cover image into the cover image in the same way as DS [10]. The batch size in training is set to 16, and each batch is divided into two parts, one of which is the face images (or face features)  $F$  and the other is the cover images  $C$ . The proposed networks are trained for 150 epochs using Adam [39] as an optimizer. The initial value of the learning rate is set to  $10^{-5}$ , and the learning rate is multiplied by 0.2 if the loss on the validation dataset does not decrease for five consecutive epochs. The weight parameters of HN and EN for the epoch with the smallest loss for the validation dataset are used in the evaluation. Using the proposed method trained on CelebA, we evaluate the quality of stego images generated by embedding face images (or face features) into cover images and the performance of face recognition using stego images.

In the experiments, we use FaceNet [4] and ArcFace [5] as face feature extractors commonly used in training networks and in evaluating face recognition performance, where both extractors have been trained on the VGGFace2 dataset<sup>6</sup>. We use an implementation based on Inception ResNet V1<sup>7</sup> for

FaceNet, and an implementation based on IResNet-50<sup>8</sup> for ArcFace. The Euclidean distance between face features is used as the matching score when FaceNet is used, and the cosine similarity is used when ArcFace is used. In the case of a face image as input, the hyperparameters of the loss function are set to  $\alpha = 2.0$  and  $\beta = 5.0$  for FaceNet and  $\alpha = 0.25$  and  $\beta = 5.0$  for ArcFace. In the case of face features as input, the hyperparameters are set to  $\alpha = 0.5$  and  $\beta = 5.0$  for FaceNet and  $\alpha = 0.5$  and  $\beta = 0.5$  for ArcFace. Data augmentation is also applied during training to flip the left and right sides of the input data with a probability of 50%.

### C. Evaluation Metrics

We use the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) to evaluate the image quality of the stego images generated by each method, which indicates the degree of image quality degradation between the cover image  $C$  and the stego image  $S$ . PSNR used in this paper is defined by

$$\text{PSNR} = -10 \log_{10} \text{MSE}, \quad (5)$$

where

$$\text{MSE} = \frac{1}{N_1 N_2 N_3} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} \sum_{n_3=1}^{N_3} [C(n_1, n_2, n_3) - S(n_1, n_2, n_3)]^2. \quad (6)$$

$(n_1, n_2)$  indicates the image coordinates,  $n_3$  indicate the color channels,  $(N_1, N_2)$  indicates the image size, and  $N_3$  indicates the number of channels, i.e.,  $N_3 = 3$  in RGB images. The higher the value of PSNR, the less degradation in image quality. SSIM used in this paper is defined as follows. Let a local image with  $w \times w$  pixels extracted from  $C(n_1, n_2, j)$  and  $S(n_1, n_2, j)$  be  $c_i^j$  and  $s_i^j$ , respectively, where  $w = 11$  in this paper. SSIM between two local images  $c_i^j$  and  $s_i^j$  can be calculated by

$$\text{SSIM}(c_i^j, s_i^j) = \frac{2(\mu_{c_i^j} \mu_{s_i^j} + p_1)(2\sigma_{c_i^j s_i^j} + p_2)}{\{(\mu_{c_i^j})^2 + (\mu_{s_i^j})^2 + p_1\} \{(\sigma_{c_i^j})^2 + (\sigma_{s_i^j})^2 + p_2\}}, \quad (7)$$

where  $\mu_{c_i^j}$  and  $\mu_{s_i^j}$  indicate the mean of  $c_i^j$  and  $s_i^j$ , respectively,  $\sigma_{c_i^j}$  and  $\sigma_{s_i^j}$  indicate the standard deviation of  $c_i^j$  and  $s_i^j$ , respectively, and  $\sigma_{c_i^j s_i^j}$  indicates the covariance between  $c_i^j$  and  $s_i^j$ .  $p_1 = (k_1 L)^2$  and  $p_2 = (k_2 L)^2$ , where  $L = 255$ ,  $k_1 = 0.01$ , and  $k_2 = 0.03$ . SSIM between  $C$  and  $S$  is obtained as the average of  $\text{SSIM}(c_i^j, s_i^j)$ , which is given by

$$\text{SSIM} = \frac{1}{MN_3} \sum_{i=1}^M \sum_{j=1}^{N_3} \text{SSIM}(c_i^j, s_i^j), \quad (8)$$

where  $M$  indicates the number of local images. The value range of SSIM is  $[0, 1]$ . Two images are identical when SSIM is 1.

We use three metrics, i.e., Accuracy, Area under the Curve (AUC), and Equal Error Rate (EER), to evaluate the accuracy

<sup>6</sup>[https://github.com/ox-vgg/vgg\\_face2](https://github.com/ox-vgg/vgg_face2)

<sup>7</sup>[https://github.com/davidsandberg/facenet/blob/master/src/models/inception\\_resnet\\_v1.py](https://github.com/davidsandberg/facenet/blob/master/src/models/inception_resnet_v1.py)

<sup>8</sup><https://github.com/nizhib/pytorch-insightface/blob/master/insightface/iresnet.py>

of face recognition. Accuracy is the ratio of correct inferences with the person or with others. AUC is the area under the curve of the Receiver Operating Characteristic (ROC) curve, which consists of False Match Rate (FMR) and True Acceptance Rate (TAR), which is calculated by  $1 - \text{False Non-Match Rate (FNMR)}$ . The closer the value of AUC to 1, the higher the recognition accuracy. EER is the error rate when FMR and FNMR are equal. The lower the value of EER, the higher the recognition accuracy.

#### D. Quality of Stego Images

In the first experiment, we evaluate the quality of stego images generated by the proposed method. We employ four versions of the proposed method with different combinations of secret images and feature extractors. The version of the proposed method is indicated by suffixes. The two suffixes following “Proposed” indicate the format of the input secret information and the face feature extractor, respectively. For the secret information, “I” indicates a face image and “F” indicates face features. For the face feature extractor used to calculate  $L_{feat}$  and embed face features, “F” indicates FaceNet and “A” indicates ArcFace. The effectiveness of the reconstruction loss  $L_{rec}$  and the perception loss  $L_{perc}$  is also evaluated through this experiment. Note that  $L_{feat}$  is indispensable loss function in the proposed method, and thus  $L_{feat}$  is used for training in any case. We compare the quality of stego images generated by DS [10] and MIAIS [11] with those generated by the proposed methods to demonstrate the effectiveness of the proposed method. We embed 100 randomly selected face images in LFW against cover image A<sup>9</sup> and cover image B<sup>10</sup>, which are shown in Fig. 9, to evaluate the quality of the stego image. Note that by fixing the seed of the random number, the same face image is selected for evaluating stego images generated by each method. Table I summarizes the quantitative evaluation of the quality of the stego images generated by each method, where the average of SSIM and PSNR are indicated. DS generates higher quality stego images in terms of both SSIM and PSNR when compared to MIAIS. In the proposed method, when FaceNet is used as the feature extractor,  $L_{perc}$  is more effective than  $L_{rec}$  in generating high-quality stego images. On the other hand, when ArcFace is used as the feature extractor, it is impossible to generate a stego image that is close to the appearance of the cover image without introducing  $L_{rec}$ . By introducing  $L_{rec}$  and  $L_{perc}$  into the proposed method, both SSIM and PSNR are significantly improved in all versions. Therefore, the use of both  $L_{rec}$  and  $L_{perc}$  as loss functions allows us to generate high-quality stego images regardless of the type of feature extractor. In the following, we use the proposed method with both  $L_{rec}$  and  $L_{perc}$  in experiments. Fig. 9 shows examples of stego images generated by each method. The stego images generated by DS have the same appearance as the cover image, but when enlarged, a grid-like

noise pattern can be seen<sup>11</sup>. The stego images generated by MIAIS show a grid-like noise pattern stronger than that of DS. On the other hand, the stego images generated by the proposed methods have the same appearance as the cover image even when enlarged, and the quality of the stego images is quite high.

If an image that can be obtained from the Internet is used as a cover image  $C$ , it may be possible to extract embedded data from the difference between the stego image  $S$  and the cover image  $C$ . To evaluate the quality of the stego image, the possibility of extracting embedded data from the stego image is qualitatively verified from the difference between the stego image and the cover image. In this experiment, a landscape image<sup>12</sup> and a black image are used as cover images  $C$ . The black image, i.e., the image with all pixel values set to 0 in the RGB image, is one of the cover images that is difficult to embed secret data. Fig. 10 shows two example pairs of the embedding face image and the cover image  $C$ , their stego images  $S$  generated by DS, MIAIS and the proposed methods, and amplified difference images between the stego image  $S$  and the cover image  $C$  by a factor of 10. In DS, the face outline appears in the difference for both cover images. In MIAIS, the face appears in the difference for both cover images. In particular, when a black image is used as the cover image, we can observe that the face clearly appears in the difference. The proposed methods do not show any embedded data in the difference when either cover image is used. We have confirmed similar trends for each method through qualitative evaluation of stego images using 100 face images of LFW. Thus, the proposed methods can safely embed the secret data even when cover images with little texture are used. We quantitatively evaluate whether face features remain in the difference image between the stego image  $S$  and the cover image  $C$ . The feature vectors are extracted from the difference image  $S - C$  and the original face image using a feature extractor, and the cosine similarity between them is calculated. In generating the stego images, we use a landscape image<sup>12</sup> and a black image as cover images as shown in Fig. 10. We select 100 images from LFW, perform the above evaluation, and summarize the average of the cosine similarity as shown in Table II. We observe that the cosine similarity is small for all the methods, and that the difference between the stego image and the cover image does not include the features of the original face image.

Steganalysis, which is an analysis technique to check whether any information is embedded in an image, is used to evaluate the quality of stego images in steganography. StegExpose<sup>13</sup>, which was also used to evaluate stego images generated by DS and MIAIS, is also used in this paper. StegExpose evaluates the score using multiple analysis methods of detecting Least Significant Bit (LSB)-based steganography, and identifies an input image as a stego image when the score

<sup>11</sup>When zooming in on the figure in the PDF viewer and focusing on the white area of the soccer ball, the cover image and the stego images generated by the proposed method are smooth, while block noise is observed in the stego image generated by DS. Also, when focusing on the nose of the airplane, the stego image generated by DS contains vertical line noise.

<sup>12</sup><https://www.photo-ac.com/main/detail/33285>

<sup>13</sup><https://github.com/b3dk7/StegExpose>

<sup>9</sup><https://www.photo-ac.com/main/detail/334941>

<sup>10</sup><https://www.photo-ac.com/main/detail/4046627>



TABLE I  
QUANTITATIVE EVALUATION OF THE QUALITY OF THE STEGO IMAGES GENERATED BY EACH METHOD

Method	Secret data	Feature extractor	$L_{rec}$	$L_{perc}$	Cover image A		Cover image B	
					SSIM $\uparrow$	PSNR [dB] $\uparrow$	SSIM $\uparrow$	PSNR [dB] $\uparrow$
DS [10]	Face image	—	—	—	0.9137	34.30	0.9551	36.14
MIAIS [11]	Face image	—	—	—	0.7245	23.74	0.8428	29.78
Proposed_I_F	Face image	FaceNet	$\checkmark$	$\checkmark$	0.9651	36.38	0.9619	34.65
					0.9905	44.15	0.9873	40.96
Proposed_I_A	Face image	ArcFace	$\checkmark$	$\checkmark$	0.9925	45.34	0.9895	41.74
					0.9188	25.18	0.9091	20.60
Proposed_F_F	Face feature	FaceNet	$\checkmark$	$\checkmark$	0.3425	7.38	0.2444	9.22
					0.9892	43.85	0.9862	40.67
Proposed_F_A	Face feature	ArcFace	$\checkmark$	$\checkmark$	0.9834	41.02	0.9784	38.59
					0.9913	44.54	0.9987	41.40
Proposed_F_A	Face feature	ArcFace	$\checkmark$	$\checkmark$	0.9916	44.62	0.9901	42.02
					0.9033	31.85	0.9271	31.86
			$\checkmark$	$\checkmark$	0.2988	7.39	0.2060	9.40
			$\checkmark$	$\checkmark$	0.9890	43.53	0.9861	40.55

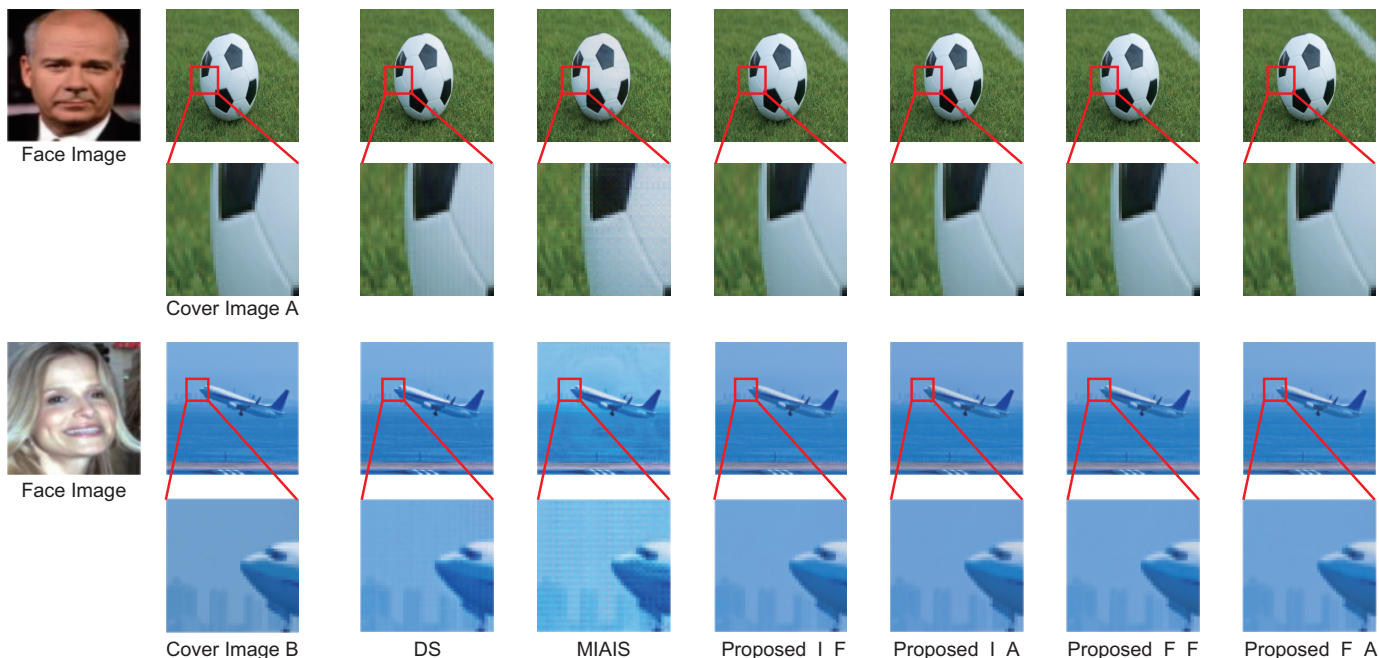


Fig. 9. Examples of stego images generated by each method.

TABLE II  
AVERAGE OF THE COSINE SIMILARITY BETWEEN THE FEATURE VECTOR EXTRACTED FROM THE DIFFERENCE IMAGE BETWEEN THE STEGO IMAGE  $S$  AND THE COVER IMAGE  $C$  AND THE FEATURE VECTOR OF THE ORIGINAL FACE IMAGE FOR 100 FACE IMAGES IN LFW.

Method	FaceNet		ArcFace	
	landscape	black	landscape	black
DS	0.1526	0.1681	-0.0031	-0.0047
MIAIS	0.0954	0.1414	0.0231	0.0011
Proposed_I_F	0.1680	0.1684	-0.0058	-0.0056
Proposed_I_A	0.1674	0.1684	-0.0056	-0.0059
Proposed_F_F	0.1677	0.1685	-0.0055	-0.0056
Proposed_F_A	0.1670	0.1684	-0.0057	-0.0055

is above a threshold. In this experiment, 1,000 cover images randomly selected from ImageNet<sup>14</sup>, the face image of a man as shown in the upper row of Fig. 9 is embedded by each of DS, MIAIS, and the proposed methods, and 1,000 stego

images generated by each method are used as input images for StegExpose. True positive rate and false positive rate are obtained by changing the threshold value for the score output from StegExpose in the range of 0 to 1 at 0.01 intervals, the Receiver Operating Characteristic (ROC) curve is plotted, and its Area Under the Curve (AUC) is used for evaluation. If AUC is close to 0.5, it indicates that the stego image is discriminated from the cover image with a probability of half, i.e., the stego image is not distinguishable from the cover image. Fig. 11 shows ROC curves, their AUCs obtained based on the scores of StegExpose, and DET curves, which are standard in the evaluation of biometric systems, for each method. AUCs of DS and MIAIS are 0.6459 and 0.6939, respectively, and hence there is a high possibility that the stego image can be identified by Steganalysis. AUCs of Proposed\_I\_F, Proposed\_I\_A, Proposed\_F\_F, and Proposed\_F\_A are 0.5433, 0.5686, 0.5414, and 0.5836, which are close to 0.5 compared with DS and MIAIS, indicating that it is difficult to distinguish

<sup>14</sup><https://www.image-net.org/download.php>

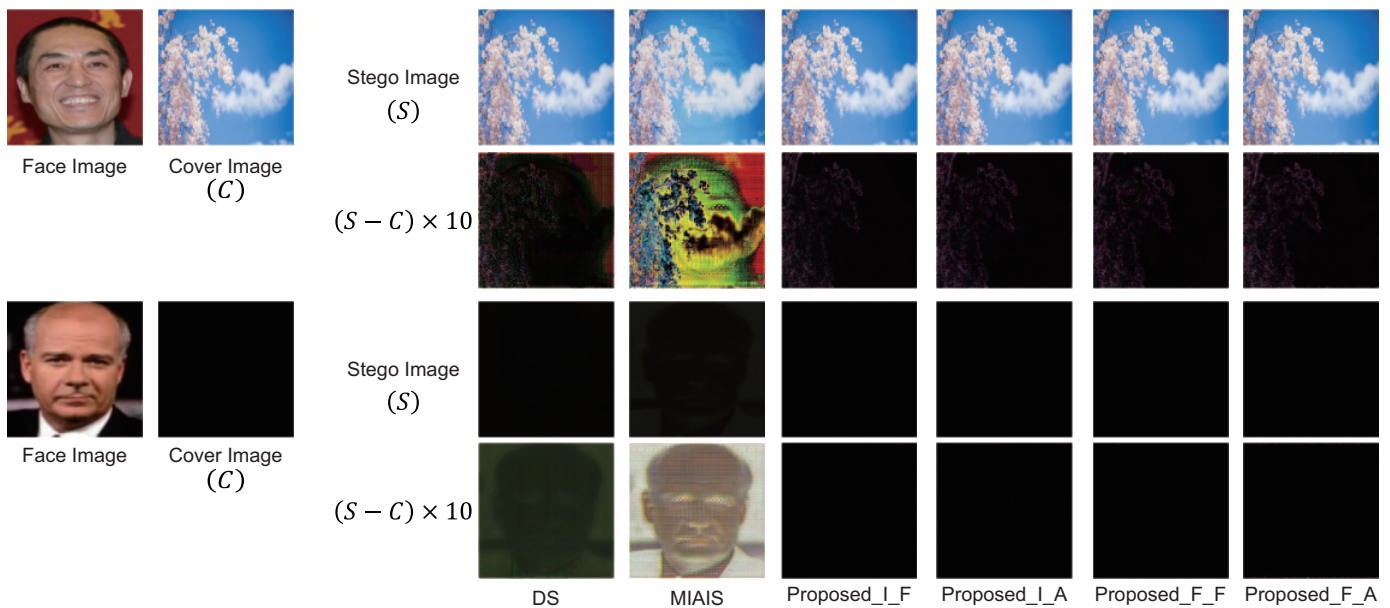






Fig. 10. Two example pairs of the embedding face image and the cover image  $C$ , their stego images  $S$  generated by DS, MIAIS and the proposed methods, and amplified difference images between the stego image  $S$  and the cover image  $C$  by a factor of 10.

TABLE III  
FACE RECOGNITION PERFORMANCE OF THE PROPOSED METHODS FOR LFW WHEN CHANGING THE TYPE OF COVER IMAGES.

Cover image	Method	Accuracy [%]	AUC	EER [%]
—	FaceNet	97.22	0.9961	2.833
—	ArcFace	99.70	0.9994	0.3000
	Proposed_I_F	95.35	0.9915	4.733
	Proposed_I_A	97.63	0.9958	2.500
	Proposed_F_F	97.13	0.9960	2.900
	Proposed_F_A	99.50	0.9991	0.6667
	Proposed_I_F	94.07	0.9874	6.000
	Proposed_I_A	97.58	0.9954	2.500
	Proposed_F_F	97.10	0.9958	3.000
	Proposed_F_A	99.50	0.9992	0.6333
	Proposed_I_F	95.43	0.9920	4.567
	Proposed_I_A	98.37	0.9976	1.767
	Proposed_F_F	97.18	0.9960	2.933
	Proposed_F_A	99.48	0.9992	0.6333
	Proposed_I_F	61.62	0.6547	38.73
	Proposed_I_A	66.77	0.7263	33.43
	Proposed_F_F	96.43	0.9935	3.700
	Proposed_F_A	99.12	0.9989	0.9667

by Steganalysis.

### E. Performance Evaluation in Face Recognition

We evaluate the recognition performance of the proposed method in face recognition to demonstrate the effectiveness of the proposed method for cancelable face recognition.

First, we evaluate the recognition performance of the proposed method using LFW as the test dataset when changing the type of the cover images  $C$ . Stego images  $S_1$  and  $S_2$  are generated by embedding each of the face image pairs (or face feature pairs)  $F_1$  and  $F_2$  from the test dataset into the cover image  $C$  using HN. Matching scores are obtained by comparing the features  $f'_{face_1}$  and  $f'_{face_2}$  extracted from the stego images  $S_1$  and  $S_2$  using EN. We evaluate the recognition performance of the proposed method by embedding LFW face

images in cover images: face<sup>15</sup>, landscape<sup>16</sup>, black, and noise images. The experiment setup except for the cover images is the same as in Sect. IV-B. Table III shows the experimental results. The first two rows of this table show the baseline recognition performance of FaceNet [4] and ArcFace [5] for LFW. When face, landscape, and black images are used as cover images, the recognition performance degrades little from the baselines. On the other hand, when the noise image is used as the cover image, the recognition performance drops significantly from the baselines when the face image is embedded as secret data. The cover images with complex textures, such as noise images, were not included in the training dataset, since face images in CelebA were used to train HN and EN as described in Sect. IV-B. Hence, the effective information for face recognition could not be embedded in the stego image. Even when a noise image is used as a cover image, the proposed method with embedded face features exhibits less degradation in recognition performance. We consider that the appearance of the feature vector image has a high affinity with the noise image. This is also because the feature vectors are duplicated to be the same size as the cover image and then embedded in the cover image, resulting in less degradation or refinement of the features extracted by EN. From the above results, we found that there is almost no degradation of recognition performance when natural images are used as cover images when face images are embedded, and almost no degradation of recognition performance when face features are embedded, regardless of the type of cover images.

Next, we compare the face recognition performance of DS and the proposed method using LFW and CASIA. As in the above experiment, we use FaceNet and ArcFace as the baseline

<sup>15</sup>[https://cdn.pixabay.com/photo/2021/07/04/19/38/woman-6387396\\_960\\_720.jpg](https://cdn.pixabay.com/photo/2021/07/04/19/38/woman-6387396_960_720.jpg)

<sup>16</sup><https://www.photo-ac.com/main/detail/227510>

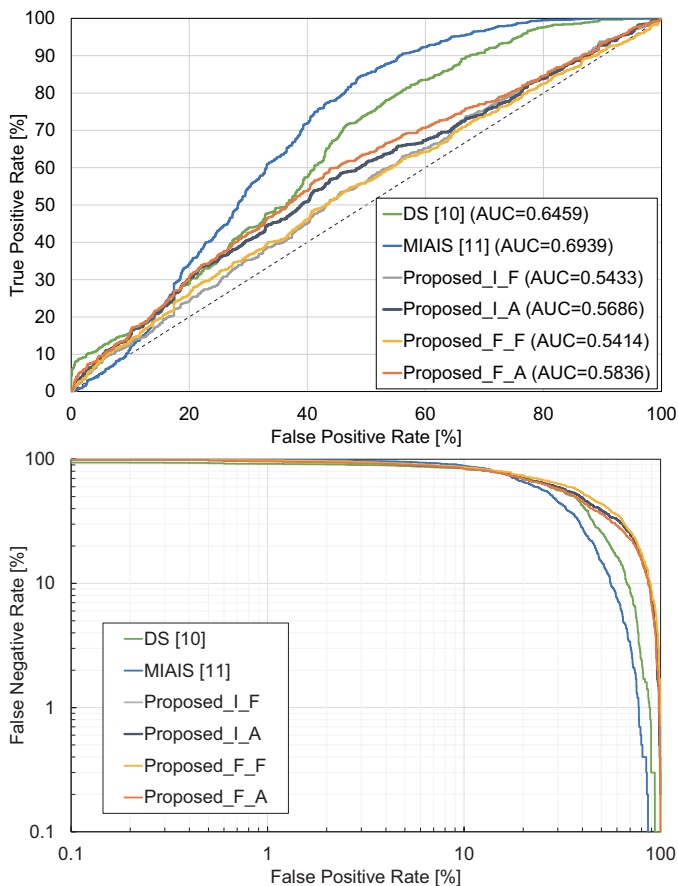


Fig. 11. ROC curves, their AUCs for each method obtained by Steganalysis with StegExpose, and DET curves.

for face recognition performance. To confirm the degradation of recognition performance due to image degradation caused by embedding and extracting images in steganography, we generate a stego image by embedding a face image into a cover image using HN of DS, and evaluate recognition performance when a face image extracted from the stego image using RN of DS is input to FaceNet or ArcFace. In this experiment, the QR code as shown in Fig. 12 is used as the cover image for DS and the proposed method, since QR codes are used as cover images in the application of the proposed method described in Sect. IV-G. Table IV summarizes the experimental results of each method for LFW and CASIA. DS exhibited slightly degraded recognition performance for both FaceNet and ArcFace. This is because the process of embedding and extracting in steganography resulted in the degradation of the face images. The proposed method extracts face features from stego images, which are generated by HN, without extracting secret information using EN. When face images are embedded by the proposed method, the recognition performance is lower than that of DS. Instead of embedding a face image in a cover image and extracting it from a stego image as in DS, the proposed method embeds features effective for face recognition obtained from the input face image into the cover image. Therefore, there may be some degradation in the output face features compared to DS. On the other hand, when face features are embedded in the proposed method,



Fig. 12. Cover image of QR code used in performance evaluation of face recognition.

TABLE IV  
FACE RECOGNITION PERFORMANCE OF EACH METHOD IN LFW AND CASIA.

Dataset	Method	Accuracy [%]	AUC	EER [%]
LFW	FaceNet [4]	97.22	0.9961	2.833
	DS [10]	96.83	0.9945	3.333
	Proposed_I_F	94.33	0.9878	5.700
	Proposed_F_F	97.17	0.9960	2.867
	ArcFace [5]	99.70	0.9994	0.3000
	DS [10]	99.42	0.9992	0.6667
CASIA	Proposed_I_A	97.78	0.9967	2.400
	Proposed_F_A	99.50	0.9991	0.6667
	FaceNet [4]	84.83	0.9160	15.96
	DS [10]	83.52	0.9066	16.97
	Proposed_I_F	78.92	0.8675	21.18
	Proposed_F_F	84.86	0.9159	15.75
CASIA	ArcFace [5]	92.83	0.9541	9.349
	DS [10]	92.59	0.9530	9.553
	Proposed_I_A	84.19	0.9065	16.69
	Proposed_F_A	92.65	0.9531	9.425

the recognition performance is comparable to FaceNet and ArcFace. As discussed in the above experiment, this is because the feature vectors are duplicated to be the same size as the cover image and then embedded in the cover image, resulting in less degradation and refinement of the features extracted by EN.

Finally, we compare the face recognition performance between MIAIS and the proposed method using AFD as in [11]. In [11], MIAIS is trained using the training dataset of AFD, and the performance of face recognition on the stego images generated by MIAIS is evaluated using the test dataset of AFD. For a fair performance comparison with MIAIS, we train HN and EN of the proposed method on the training dataset of AFD to generate stego images under the same conditions as MIAIS. The face images in LFW are used as cover images in the training of the proposed method. Since ArcFace exhibited higher recognition performance than FaceNet based on the previous experiments, we use the proposed methods (Proposed\_I\_A and Proposed\_F\_A) with ArcFace as the feature extractor in this experiment. The other conditions are the same as in Sect. IV-B. The performance evaluation of face recognition uses the test dataset of AFD as in [11]. The experimental results of ArcFace, MIAIS, and the proposed methods are shown in Table V. MIAIS showed a slightly lower recognition performance than ArcFace, Proposed\_I\_A showed a significantly lower recognition performance than MIAIS, and Proposed\_F\_A showed a higher recognition performance than ArcFace. Proposed\_I\_A did not perform well due to the small number of images in the training dataset of AFD

TABLE V  
FACE RECOGNITION PERFORMANCE OF ARCFACE, MIAIS, AND THE PROPOSED METHODS IN AFD.

Method	Accuracy [%]	AUC	EER [%]
ArcFace [5]	87.80	0.9433	12.40
MIAIS [11]	86.14	0.9380	13.41
Proposed_I_A	61.65	0.6624	38.50
Proposed_F_A	92.03	0.9722	8.375

compared to the training dataset of CelebA . On the other hand, Proposed\_F\_A exhibited higher recognition performance than MIAIS since it could be sufficiently trained even when the number of images in the training dataset was small.

We summarize the results of the above experiments. The proposed method can produce higher quality stego images than DS and MIAIS. When face images are embedded in the proposed method, the quality of the stego images is higher, while the recognition performance is lower than that of DS and MIAIS. Therefore, when embedding face images, effective information for face recognition is not sufficiently embedded in the stego images. On the other hand, when embedding face features in the proposed method, we confirmed that the recognition performance is equivalent to that of the original face recognition method. Therefore, the above results demonstrate that the proposed method of embedding features is effective from the viewpoint of cancelable face recognition using stego images.

### F. Security Analysis

We verify that the proposed method satisfies non-invertibility by training EN to extract the feature vector agitated by the pseudo-random matrix  $R$  at feature loss  $L_{feat}$  as shown in Fig. 6 (c). In the performance evaluation of face recognition as shown in Table IV, both  $f_r$  and  $f_i$  are feature vectors agitated by  $R$  since stego images are generated for both registration and authentication phases as shown in Fig. 3. To verify the effectiveness of  $R$ , we evaluate the performance of face recognition using the feature vector obtained by the feature extractor on one of the pair and the feature vector obtained from the stego image using EN on the other. For genuine pairs of LFW, one of the images is selected, and the cosine similarity between the feature vector obtained from the feature extractor and the feature vector obtained from its stego image using EN is calculated. Since the cosine similarity of the same image is obtained, the degradation of the feature vectors by the proposed method can be confirmed. For impostor pairs of LFW, a feature vector is extracted from one of the images using a feature extractor and a feature vector is extracted from the stego image of the other using EN and the cosine similarity between the feature vectors is calculated. In this experiment, ArcFace is used as the feature extractor, and face image<sup>15</sup> and landscape image<sup>16</sup> are used as cover images. Table VI shows the results when  $R$  is not used in training EN, and Table VII shows the results when  $R$  is used in training EN. When the pseudo-random matrix  $R$  is not used, the features of the original face image can be extracted from the stego image, and thus almost perfect authentication can be achieved. On the other hand, when EN is trained to output a feature vector that is

TABLE VI  
FACE RECOGNITION PERFORMANCE WITH FEATURE VECTORS EXTRACTED USING EN TRAINED *without* PSEUDO-RANDOM MATRIX  $R$ .





Cover image	Method	Accuracy [%]	AUC	EER [%]
	Proposed_I_F	99.83	0.9999	0.200
	Proposed_I_A	99.80	0.9999	0.2333
	Proposed_F_F	100	1.0	0.0
	Proposed_I_A	100	1.0	0.0
	Proposed_I_F	98.72	0.9991	1.267
	Proposed_I_A	99.77	0.9999	0.233
	Proposed_F_F	100	1.0	0.0
	Proposed_F_A	100	1.0	0.0

TABLE VII  
FACE RECOGNITION PERFORMANCE WITH FEATURE VECTORS EXTRACTED USING EN TRAINED *with* PSEUDO-RANDOM MATRIX  $R$ .

Cover image	Method	Accuracy [%]	AUC	EER [%]
	Proposed_I_F	50.00	0.3991	56.63
	Proposed_I_A	50.00	0.3897	58.47
	Proposed_F_F	50.03	0.3839	58.30
	Proposed_I_A	50.02	0.3351	62.27
	Proposed_I_F	50.00	0.4245	55.50
	Proposed_I_A	50.00	0.3879	58.40
	Proposed_F_F	50.18	0.4170	58.37
	Proposed_F_A	50.02	0.3351	56.17

agitated using  $R$ , it is almost impossible to authenticate since the feature vector obtained from the stego image is different from the one extracted from the original face image. From the above, we can verify that the proposed method can embed the feature vectors obtained by the feature extractor into the stego image with almost no degradation, and that the non-invertibility is satisfied by training EN to output feature vectors agitated using  $R$  at feature loss  $L_{feat}$ .

The proposed method as shown in Fig. 3 requires that EN is securely protected, which is an indispensable condition for cancelable biometrics. The diversity is satisfied by changing the pseudo-random matrix  $R$  in the feature loss  $L_{feat}$ , since different feature vectors are output from EN even when the same face image is embedded in the stego image. The unlinkability is satisfied by training EN with different  $R$  for different applications. If a malicious third party can easily access EN, he/she can use EN to extract feature vectors from stego images, which are agitated by a pseudo-random matrix, and thus may estimate the pseudo-random matrix or recover embedded face images. In such a case, the revocability is satisfied since HN and EN can be retrained by changing the pseudo-random matrix  $R$ , thereby discarding the stego images generated by the previous HN. Compared to typical template protection approaches that require users to manage their own user-specific parameters, in our approach, EN and feature vectors extracted by EN must be managed on the server. The necessity of managing feature vectors extracted by EN on the server is discussed later. The typical approaches have a risk of leaking user-specific parameters, while our approach has the cost of re-training EN to cancel the stego images (templates) and the cost of managing EN and feature vectors extracted by EN on the server. Therefore, our approach involves little security risk to the user if EN and feature vectors extracted by EN are managed securely and appropriately on the server. In our approach, if the templates have to be canceled due to

TABLE VIII  
FACE RECOGNITION PERFORMANCE USING THE ESTIMATED  
PSEUDO-RANDOM MATRIX  $R'$ .

Pair	Accuracy [%]	AUC	EER [%]
$f'_{face}$ and $f_{face}$	99.48	0.9993	0.6333
$fR'$ and $f_{face}$	89.98	0.9631	10.20

EN leakage, it is necessary to discard all the stego images, retrain EN, and regenerate the stego images, which may result in a high cost. On the other hand, if EN and feature vectors extracted by EN can be properly and securely managed on the server, our approach is highly convenient, since users do not need to manage any parameters to ensure safety.

In the following, we discuss the worst-case scenario where HN, EN, and the feature extractor used for training are accessible. We verify whether we can estimate the pseudo-random matrix  $R$  used in the feature loss  $L_{feat}$ . We also verify whether the face image embedded in the cover image  $C$  can be reconstructed from the feature vector  $f$  extracted by EN.

First, we verify whether the pseudo-random matrix  $R$  used in training the feature loss  $L_{feat}$  can be estimated. As shown in Fig. 5, the feature vector output by EN is of 512 dimensions, and therefore the pseudo-random matrix used in the proposed method is  $R \in \mathbb{R}^{512 \times 512}$ . We need at least 512 feature vectors to estimate  $R$ . Let  $X \in \mathbb{R}^{512 \times 512}$  be the matrix of feature vectors  $f$  extracted from 512 face images using a face feature extractor and concatenated.  $Y$  is obtained by agitating  $X$  with the pseudo-random matrix  $R$  as

$$Y = XR. \quad (9)$$

Let  $Y' \in \mathbb{R}^{512 \times 512}$  be the matrix that concatenates the feature vector  $f'_{face}$  obtained by generating stego images from 512 face images and arbitrary cover images using HN and inputting them to EN. As shown in Fig. 6 (c),  $Y$  and  $Y'$  are trained to be equal, and hence this can be represented by

$$Y' \approx XR. \quad (10)$$

If the inverse of  $X$  is multiplied on both sides from the left, we obtain

$$X^{-1}Y' \approx X^{-1}XR = R. \quad (11)$$

From the above, it is possible to estimate  $R$  from  $X$  and  $Y'$ . We evaluate the accuracy of the estimated pseudo-random matrix  $R' = X^{-1}Y'$  through the experiment on face recognition. In this experiment, we use LFW as the face image dataset, ArcFace as the feature extractor, Proposed\_FA as HN and EN, and face<sup>15</sup> as the cover image. We perform face recognition using feature  $fR'$ , which is the face feature  $f$  output from the feature extractor and agitated by  $R'$  and feature  $f'_{face}$  extracted from a stego image using EN. Table VIII shows the results when  $f'_{face}$  pairs are used and when one of the pairs is  $fR'$ . Although  $R$  is not completely estimated since the recognition accuracy is reduced by using  $R'$ , it can be estimated to a certain level. Therefore, it is essential to securely manage EN and protect the feature vectors output from EN.

Next, we verify whether face images can be recovered from the feature vectors output by EN and whether they can

be used for face recognition. In this paper, we use NbNet [40], which is a method for recovering face images from face feature vectors. We train NbNet to recover face images embedded in stego images from features extracted from stego images using EN. [40] uses face images generated by DCGAN [41] for training, while this experiment uses high-resolution face images generated by StyleGAN2 [42]. A stego image is obtained by embedding the generated face image into the cover image using HN, and the face feature vector extracted from the stego image using EN is used as the input to NbNet. Other experimental conditions are the same as for [40]. Since the image size output from NbNet is  $160 \times 160$  pixels, the face feature extractor and the proposed method are also trained on images of  $160 \times 160$  pixels. Note that the performance is lower than that of Table IV since ArcFace trained with an image size of  $160 \times 160$  pixels is used as the face feature extractor. Fig. 13 shows the results of face image reconstruction from the feature vectors extracted from EN of the proposed method and the cosine similarity between the original face image and the reconstructed face image. When a face image is embedded, a face image different from the original face image is reconstructed, while when face features are embedded, face attributes such as gender, beard, and glasses are reconstructed. Table IX shows the face recognition performance when only the original face image is used and when one of the matching pairs is replaced by the recovered image. The recognition accuracy is significantly lower when the face image is embedded with HN, while the recognition accuracy is also slightly lower when the face features are embedded with HN. As shown in Table IV, the proposed method with embedded features has less degradation in recognition performance, and this is because features effective for face recognition are embedded in the stego image. If the feature vector can be extracted from the stego image using EN, there is a possibility that the embedded face image can be recovered from the stego image. Therefore, it is essential to securely protect EN. Furthermore, it is possible to limit the generation of the stego images using HN, and to generate the stego images on the server side instead of distributing them to the user's device. As shown in Table IV, the proposed method with embedded features has less degradation in recognition performance, and this is because features effective for face recognition are embedded in the stego image. If the feature vector can be extracted from the stego image using EN, there is a possibility that the embedded face image can be recovered from the stego image. Therefore, it is essential to securely protect EN. Furthermore, we can consider countermeasures to limit the generation of the stego images using HN, and to generate the stego images on the server side instead of generating them on the user's devices.

Finally, we quantitatively evaluate the unlinkability of the steganography-based methods according to [16]. To evaluate unlinkability, we obtain the mated and non-mated sample distributions. To obtain the mated sample distribution, templates  $T_1$  and  $T_2$  are created from biometric sample  $M_1$  using secret keys  $K_1$  and  $K_2$  as follows:

$$T_1 = PIE(M_1, K_1), \quad T_2 = PIE(M_1, K_2). \quad (12)$$

The secret key required by the template protection methods

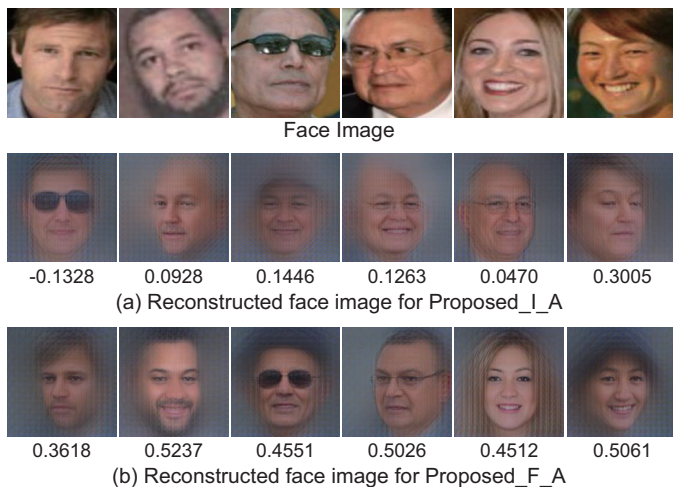


Fig. 13. Reconstructed face images from features using NbNet [40] and cosine similarity with original face images: (a) Reconstructed face images for Proposed\_I\_A and (b) Proposed\_F\_A.

TABLE IX  
FACE RECOGNITION PERFORMANCE USING RECONSTRUCTED FACE IMAGES FROM FEATURES, WHERE METHODS INDICATE THE USE OF ORIGINAL FACE IMAGES, RECONSTRUCTED IMAGES FOR PROPOSED\_I\_A AND PROPOSED\_F\_A.

Method	Accuracy [%]	AUC	EER [%]
Original	99.22	0.9987	0.9333
Proposed_I_A	54.17	0.5534	46.40
Proposed_F_A	89.10	0.9573	11.23

discussed in [16] corresponds to the cover image in the steganography-based method. A Pseudonymous Identifier Encoder (PIE) used for template generation in [16] is replaced by steganography-based methods, and the generated stego images are used as templates. The linkage score  $s$  is calculated using the linkage function  $LS$  as follows

$$s = LS(\mathbf{T}_1, \mathbf{T}_2). \quad (13)$$

In this experiment, Mean Squared Error (MSE) and Normalized Cross Correlation (NCC) are used to evaluate the similarity between stego images as  $LS$ . Similarly, to obtain the non-mated sample distribution, two template pairs are generated as

$$\mathbf{T}_1 = PIE(\mathbf{M}_1, \mathbf{K}_1), \quad \mathbf{T}_2 = PIE(\mathbf{M}_2, \mathbf{K}_2). \quad (14)$$

The score distribution is input to the public code<sup>17</sup> used in [16] for obtaining a quantitative evaluation measure of unlinkability,  $D_{\leftrightarrow}^{sys}$ . Table X summarizes the evaluation measure  $D_{\leftrightarrow}^{sys}$  for analyzing the unlinkability performance of each method. If  $D_{\leftrightarrow}^{sys}$  is less than 0.07, the method is almost fully unlinkable [16]. Since  $D_{\leftrightarrow}^{sys}$  for all the methods is less than or equal to 0.07, they can be considered almost fully unlinkable.

### G. Application

One of the potential applications of the proposed method to enhance the security of cancelable face recognition is the

TABLE X  
SUMMARY OF THE EVALUATION MEASURE  $D_{\leftrightarrow}^{sys}$  FOR ANALYZING THE UNLINKABILITY PERFORMANCE FOR EACH METHOD. IF  $D_{\leftrightarrow}^{sys}$  IS LESS THAN 0.07, THE METHOD IS ALMOST FULLY UNLINKABLE [16].

$LS$	MSE	NCC
DS [10]	0.0501	0.0502
MIAIS [11]	0.0512	0.0455
Proposed_I_F	0.0464	0.0596
Proposed_I_A	0.0528	0.0468
Proposed_F_F	0.0538	0.0460
Proposed_F_A	0.0538	0.0460

combination with one-time passwords (OTPs) using QR codes. Since the proposed method can use any image as a cover image, a QR code can be used as a cover image as shown in the second experiment in Sect. IV-E. By using this fact, we can develop the authentication architecture as shown in Fig. 14. At the registration phase, a stego image is generated using an arbitrary image as the cover image  $C$ , and then it is transferred to the authentication server. At the authentication phase, a stego image is generated using the QR code generated from OTP sent from the authentication server as the cover image. By setting the expiration time of the QR code by OTP, we prevent unauthorized use of the QR code by a third party. The stego image is transferred to the authentication server, and the server checks whether the original OTP and the returned OTP are identical. If the OTPs do not match, the authentication fails as an attack by a third party or an incorrect input. If the OTP matches, the features extracted from the transferred stego image using EN are compared with the features extracted from the stego image stored in the database. This approach can prevent attacks such that insert fake stego images into devices, communication paths, and servers without the need for supplementary information such as PINs or passwords, and can improve diversity and revocability, which are requirements for cancelable biometrics.

### V. CONCLUSION

In this paper, we proposed a cancelable face recognition method using Deep Steganography (DS), which consists of Hiding Network (HN) and Extracting Network (EN). In HN, secret information is embedded in the cover image to generate a stego image, and EN extracts face features from the stego image. Appearance of the stego image is indistinguishable from that of the cover image, and face features can be extracted from the stego image using only the dedicated EN. We introduced three loss functions: the reconstruction loss, perceptual loss, and feature reconstruction loss, to train the proposed network. Among the three loss functions, the perceptual loss contributes to generate the high-quality stego images, whose quality is higher than that of DS [10] and MIAIS [11]. We demonstrated the effectiveness of the proposed method compared to DS and MIAIS through performance and security evaluation experiments using LFW, CASIA, and AFD. We also presented one of the potential applications of the proposed method when using the QR code image as the cover image. The QR code image generated from OTP can be used as the cover image of the proposed method to perform two-factor authentication, which can improve the security without

<sup>17</sup><https://github.com/dasec/unlinkability-metric>

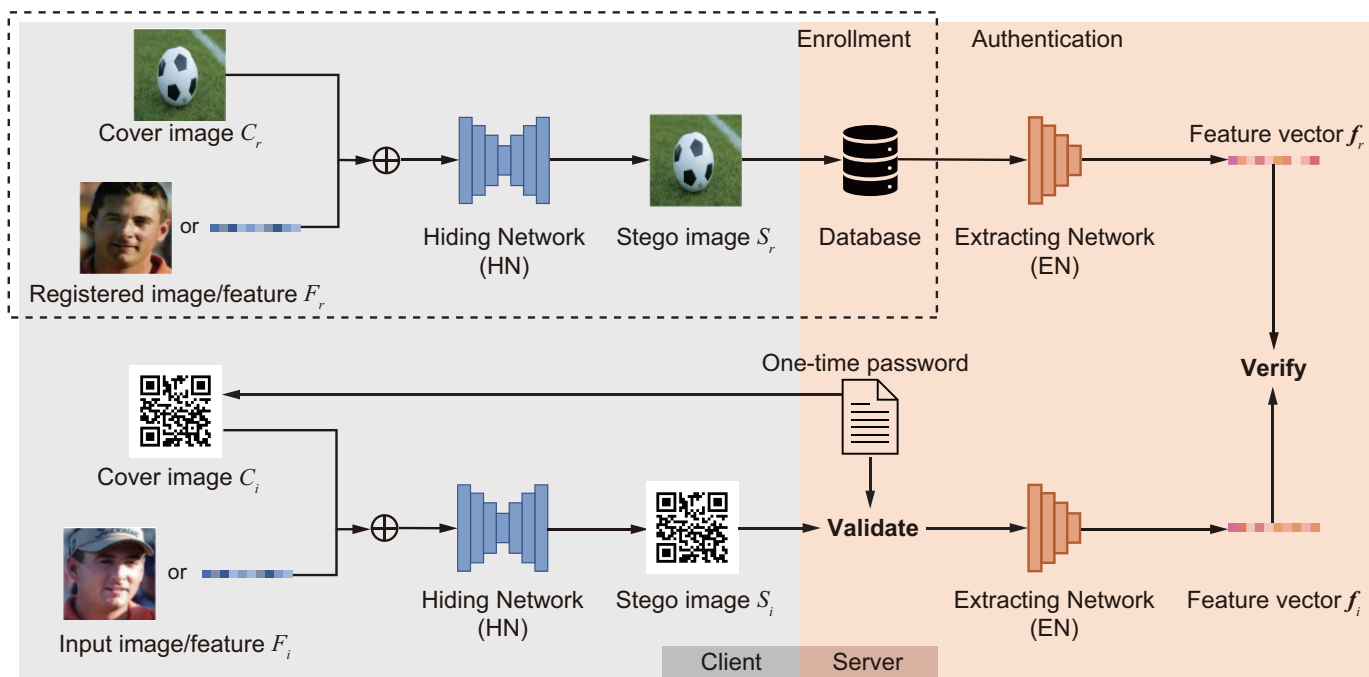


Fig. 14. Authentication architecture of cancelable face recognition using the combination of the proposed method and the QR code of a one-time password.

sacrificing the usability. We believe that this approach of using DS could be a new baseline for cancelable biometrics. In our future work, we will demonstrate the effectiveness of this approach for other biometric characteristics.

#### ACKNOWLEDGMENT

This work was supported, in part, by JSPS KAKENHI Grant Number 21H03457 and 21J15252.

#### REFERENCES

- [1] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Springer, 2008.
- [2] S. Li and A. Jain, *Handbook of Face Recognition*. Springer, 2011.
- [3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1701–1708, Jun. 2014.
- [4] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 815–823, Jun. 2015.
- [5] J. Deng, J. Guo, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 4685–4694, Jun. 2019.
- [6] C. Soutar, D. Roberge, A. Stoianov, R. M. Gilroy, and B. V. V. Kumar, “Biometric encryption using image processing,” *Electronic Imaging*, pp. 1–11, 1 1998.
- [7] M. Rawat and N. Kumar, “Cancelable biometrics: A comprehensive survey,” *Artificial Intelligence Review*, vol. 53, p. 3403–3446, 6 2020.
- [8] B. Choudhury, P. Then, V. Raman, B. Issac, and M. K. Haldar, “Cancelable iris biometrics based on data hiding schemes,” *2016 IEEE Student Conference on Research and Development (SCoReD)*, pp. 1–6, Dec. 2016.
- [9] G. J. Simmons, “The prisoners’ problem and the subliminal channel,” *Advances in Cryptology (Proc. CRYPTO ’83)*, pp. 51–67, Jun. 1983.
- [10] S. Baluja, “Hiding images in plain sight: Deep steganography,” *Proc. Advances in Neural Information Processing Systems*, vol. 30, pp. 2069–2079, Dec. 2017.
- [11] J. Cui, P. Zhang, S. Li, L. Zheng, C. Bao, J. Xia, and X. Li, “Multitask identity-aware image steganography via minimax optimization,” *IEEE Trans. Image Processing*, vol. 30, pp. 8567–8579, Sep. 2021.

- [12] J. Johnson, A. Alahi, and F. Li, “Perceptual losses for real-time style transfer and super-resolution,” *Proc. European Conf. Computer Vision*, pp. 694–711, Mar. 2016.
- [13] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *CoRR*, vol. abs/1411.7923, pp. 1–9, Nov. 2014.
- [14] Z. Xiong, Z. Wang, C. Du, R. Zhu, J. Xiao, and T. Lu, “An Asian face dataset and how race influences face recognition,” *Lecture Notes in Computer Science (Proc. Pacific Rim Conf. Multimedia)*, vol. 11165, pp. 372–383, Sep. 2018.
- [15] A. Jain, K. Nandakumar, and A. Nagar, “Biometric template security,” *EURASIP J. Advances in Signal Processing*, vol. 2008, pp. 1–17, Mar. 2008.
- [16] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, “General framework to evaluate unlinkability in biometric template protection systems,” *IEEE Trans. Information Forensics and Security*, vol. 12, no. 6, Jun. 2018.
- [17] N. Ratha, J. Connell, and R. Bolle, “Enhancing security and privacy in biometrics-based authentication systems,” *IBM Systems J.*, vol. 40, pp. 614–634, Jan. 2001.
- [18] H. Kaur and P. Khanna, “Biometric template protection using cancelable biometrics and visual cryptography techniques,” *Multimedia Tools and Applications*, vol. 75, pp. 16333–16361, Dec. 2016.
- [19] A. Teoh, D. Ngo, and A. Goh, “Biohashing: two factor authentication featuring fingerprint data and tokenised random number,” *Pattern Recognition*, vol. 37, no. 11, pp. 2245–2255, Aug. 2004.
- [20] M. Savvides, B. Vijaya Kumar, and P. Khosla, “Cancelable biometric filters for face recognition,” *Proc. Int’l Conf. Pattern Recognition*, vol. 3, pp. 922–925, Aug. 2004.
- [21] N. F. Johnson and S. Jajodia, “Exploring steganography: Seeing the unseen,” *Computer*, vol. 31, no. 2, pp. 26–34, Feb. 1998.
- [22] D.-A. Wu and W.-H. Tsai, “A steganographic method for images by pixel-value differencing,” *Pattern Recognition Letters*, vol. 24, no. 9–10, pp. 1613–1626, Jun. 2003.
- [23] M. T. Parvez and A. A.-A. Gutub, “RGB intensity based variable-bits image steganography,” *Proc. IEEE Asia-Pacific Services Computing Conf.*, pp. 1322–1327, Jan. 2008.
- [24] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, “Techniques for data hiding,” *IBM Systems J.*, vol. 35, no. 3.4, pp. 313–336, Nov. 1996.
- [25] M. Chaumont and W. Puech, “A DCT-based data-hiding method to embed the color information in a JPEG grey level image,” *Proc. European Signal Processing Conf.*, pp. 1–5, Sep. 2006.
- [26] K. S. Babu, K. S. Raja, K. K. Kiran, T. H. Manjula-Devi, K. R. Venugopal, and L. M. Patnaik, “Authentication of secret information

in image steganography,” *Proc. IEEE Region 10 Conf.*, pp. 1–6, Dec. 2008.

[27] D. Volkhonskiy, I. Nazarov, B. Borisenko, and E. Burnaev, “Steganographic generative adversarial networks,” *Proc. NIPS 2016 Workshop on Adversarial Training*, pp. 1–8, Mar. 2016.

[28] J. Hayes and G. Danezis, “Generating steganographic images via adversarial training,” *Proc. Advances in Neural Information Processing Systems*, vol. 30, pp. 1951–1960, Dec. 2017.

[29] A. Rehman, R. Rahim, M. Nadeem, and S. Hussain, “End-to-end trained CNN encode-decoder networks for image steganography,” *Proc. European Conf. Computer Vision*, pp. 723–729, Nov. 2017.

[30] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1386–1393, Jun. 2014.

[31] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” *Proc. Int’l Conf. Medical Image Computing and Computer Assisted Intervention*, pp. 234–241, Oct. 2015.

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770–778, Jun. 2016.

[33] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 7132–7141, Jun. 2018.

[34] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.

[35] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” *Proc. Int’l Conf. Computer Vision*, pp. 3730–3738, Dec. 2015.

[36] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07–49, Oct. 2007.

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *CoRR*, vol. abs/1411.7923, pp. 1–9, Nov. 2014.

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multi-task cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Apr. 2016.

[39] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *Proc. Int’l Conf. Learning Representations*, vol. abs/1412.6980, pp. 1–15, May 2015.

[40] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, “On the reconstruction of face images from deep face templates,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 5, May 2019.

[41] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *CoRR*, vol. abs/1511.06434, pp. 1–16, Nov. 2016.

[42] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 8107–8116, Jun. 2020.



**Koichi Ito** received the B.E. degree in electronic engineering, and the M.S. and Ph.D. degree in information sciences from Tohoku University, Sendai, Japan, in 2000, 2002 and 2005, respectively. He is currently an Associate Professor of the Graduate School of Information Sciences at Tohoku University. From 2004 to 2005, he was a Research Fellow of the Japan Society for the Promotion of Science. His research interest includes signal and image processing, computer vision, and biometric authentication.



**Takashi Koza** received the B.E. degree in information engineering, and the M.S. degree in information sciences from Tohoku University, Sendai, Japan, in 2020 and 2022, respectively. He is currently working on Amazon Web Service, Japan. His research interest includes computer security and biometric authentication.



**Hiroya Kawai** received the B.E. degree in communication engineering, and the M.S. and Ph.D. degree in information sciences from Tohoku University, Sendai, Japan, in 2018, 2020, and 2023, respectively. He is currently working on Biometrics Research Laboratories, NEC Corporation, Japan. From 2021 to 2023, he was a Research Fellow of the Japan Society for the Promotion of Science. His research interest includes machine learning and biometric authentication.



**Goki Hanawa** received the B.E. degree in information engineering from Tohoku University, Sendai, Japan, in 2022. He is currently a master course student of the Graduate School of Information Sciences at Tohoku University. His research interest includes computer security and biometric authentication.



**Takafumi Aoki** received the BE, ME, and DE degrees in electronic engineering from Tohoku University, Sendai, Japan, in 1988, 1990, and 1992, respectively. He is currently a professor in the Graduate School of Information Sciences (GSIS) at Tohoku University. Since April 2018, he has also served as the Executive Vice President of Tohoku University. His research interests include theoretical aspects of computation, computer design and organization, LSI systems for embedded applications, digital signal processing, computer vision, image processing, biometric authentication, and security issues in computer systems. He received more than 20 academic awards as well as distinguished service awards for his contributions to victim identification in the 2011 Great East Japan Disaster.