IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 May 2022, accepted 12 June 2022, date of publication 21 June 2022, date of current version 5 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3185121

Detection Enhancement for Various Deepfake Types Based on Residual Noise and Manipulation Traces

JIHYEON KANG[1,2], SANG-KEUN JI[3], SANGYEONG LEE[4], DAEHEE JANG[5], AND JONG-UK HOU[4]

[1]Graduate School of Information Security, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
[2]Webtoon AI, NAVER WEBTOON Corporation, Seongnam 13529, South Korea
[3]School of Computing, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
[4]School of Software, Hallym University, Chuncheon 24252, Republic of Korea
[5]Department of Security Engineering, Sungshin Women's University, Seoul 02844, Republic of Korea

Corresponding author: Jong-Uk Hou (juhou@hallym.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIT) under Grant NRF-2020R1C1C1013433, and in part by the Hallym University Research Fund 2021 under Grant HRF-202111-003.

ABSTRACT As deepfake techniques become more sophisticated, the demand for fake facial image detection continues to increase. Various deepfake detection techniques have been introduced but detecting all types of deepfake images with a single model remains challenging. We propose a technique for detecting various types of deepfake images using three common traces generated by deepfakes: residual noise, warping artifacts, and blur effects. We adopted a network designed for steganalysis to detect pixel-wise residual-noise traces. We also consider landmarks, which are the primary parts of the face where unnatural deformations often occur in deepfake images, to capture high-level features. Finally, because the effect of a deepfake is similar to that of blurring, we apply features from various image quality measurement tools that can capture traces of blurring. The results demonstrate that each detection strategy is efficient, and that the performance of the proposed network is stable and superior to that of existing detection networks on datasets of various deepfake types.

INDEX TERMS Deepfake forensics, image forensics, residual noise, warping artifact, image quality measurement.

I. INTRODUCTION

Deepfake is a technique for creating synthetic content by naturally changing the human face of the original content using an autoencoder and generative adversarial network (GAN) [1]–[3]. In a broad sense, deepfake refers to deformed or created content that uses deep learning methods (audio deepfake [4], imaginary people generation [5], *etc.*) to trick people. In a narrower sense, deepfake refers to an image or video of a human face that has been generated using deep learning methods and can cause malicious effects. Deepfake in this narrow sense can be classified into three types according to the type of manipulation: face-swap, puppet-master, and attribute-change.

In face-swap deepfakes [6], [7], which is the most common type, a person's face is pasted onto that of another person, while maintaining the original person's expression. Previous face-swap methods consider only the shape, direction, and skin color of the face, regardless of the original facial expression. However, in deepfake, face swapping is synthesized by imitating the expression. Nowadays, because anyone can easily use deepfake face-swap methods on the Internet, it is already being exploited to naturally synthesize the face of a famous celebrity into pornography. Only celebrities have been targeted, because deepfake training requires many pictures of the same person. However, deepfake synthesis techniques that use only a few photos have been proposed; thus, anyone on the street can now be a victim of deepfake. Currently, most of the harmful effects of deepfakes fall under this category in which personal rights can be violated. Puppet-master deepfakes [8], [9], also called reenactment, manipulate a target image to follow the movements of the face, head, and upper body of a source image. As this type of deepfake does not require the appearance of other faces,

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

a more sophisticated synthesis is possible. This technique is primarily used to create fake news; fake news that synthesizes the face of a key person (president, prime minister, famous news anchor, etc.) that can cause mass confusion in society. Finally, attribute-change techniques [10], [11] can manipulate a wide range of visual traits in facial images (hair color, beard, aging signs, *etc.*). This type of deepfake can be exploited by manipulating evidence, such as changing the facial traits of a criminal captured on camera, causing social confusion. Many studies are currently in progress to detect deep fakes that can adversely affect society, but more sophisticated deepfake creation and detection avoidance methods are emerging [12], [13]. A competitive race between more elaborate deepfake generations and more accurate detection is underway.

Deepfake images differ in the form and degree of traces because of the diversity of generating algorithms, facial characteristics, and postprocessing methods. However, some tracing forms are commonly observed: generating fine noise while passing through a GAN or auto-encoder, blurring caused by resizing and postprocessing, and warping caused by a failure of facial geometric and illuminance predictions.

In this study, we propose a generalized detection method using traces to detect three types of deepfake (face-swap, puppet-master, and attribute-change). To improve detection performance, we developed a network based on image quality measurement (IQM) features and warping artifacts extracted from facial landmarks. Instead of using a general network of recent algorithms, such as XceptionNet [14], we propose the use of a network designed for steganalysis to capture residual noise traces in deepfake images. The experiments were performed using different types of deepfakes with public databases, and we demonstrate that the proposed network achieves performance stability and is superior to existing detection networks on datasets of various deepfake types.

Our contributions are summarized as follows:

- We propose a generalized detection method using traces to detect three types of deepfake: face swap, puppet-master, and attribute change.
- We developed a network based on image quality measurement (IQM) features and warping artifacts extracted from facial landmarks.
- We propose using a network designed for steganalysis to capture residual noise traces in deepfake images.

II. RELATED WORK

A. DEEPFAKE GENERATION METHODS

1) FACE-SWAP

Korshunova *et al.* [15] suggested training a multi-scale architecture convolutional neural network (CNN) to paste faces from one image to another. RSGAN [16] performs natural face swapping by separating the hair and face in a latent space. In addition, in combination with existing 3D analysis technology, the face-swap has become more sophisticated [17], [18]. Li *et al.* [6] suggested techniques (mask area adjustment, additional layer in auto-encoder, and effective post-processing) to obtain better quality face-swap content. Using these techniques, they built the 'Celeb-DF' deepfake dataset. The DeepFaceLab team [7] released the deepfake face-swap application. They used a GAN with an autoencoder and set of attention masks to improve the details of output images.

2) PUPPET-MASTER

Suwajanakorn *et al.* [8] proposed a synthesis technique for manipulating lip shape. By learning the mapping of audio features to mouth shapes, they created a fake version of a video of a Barack Obama speech using the target audio. Tripathy *et al.* [9] proposed a two-stage GAN using a facial attribute vector consisting of the head pose and action unit (AU). This model generates a neutral image with a central pose and neutral expression from a source image and transforms it to follow the target image's attributes. Rössler *et al.* [19] introduced a face manipulation dataset generated by Face2Face [20], which is a technique for facial reenactment manipulation.

3) ATTRIBUTE-CHANGE

Choi *et al.* [10] proposed StarGAN, which uses only a single model for multiple attribute domains, and Pumarola *et al.* [21] used facial AU labels that allow the generation of detailed and continuous facial expression transformations. Kingma and Dhariwal [11] proposed Glow, which uses a flow-based generative model using invertible 1×1 convolution. Glow allows various attribute changes and exhibits a high quality.

B. DEEPFAKE DETECTION METHODS

In addition to the image forensics of general image modifications [22]–[24], several deepfake detection techniques have been proposed [25]–[28]. Matern *et al.* [29] used color mismatch in two eyes and noise owing to inaccurate geometric predictions and inaccurate light predictions. Yan *et al.* [30] detected a deepfake using the inconsistency in the 3D direction between the narrow face area and overall head. Afchar *et al.* [25] proposed two simple fake-face detection networks (Meso-4 and MesoInception4) that exploit mesoscopic features. Because training is performed with a distribution in the RGB color space, Li *et al.* [31] changed the color space to HSV or YCbCr and detected deepfake using the statistical difference between the color spaces. Koopman *et al.* [32] detected deepfake videos using photo response non-uniform noise patterns that disappeared when the facial area was modified. Li *et al.* [33] identified deepfakes by representing the blending boundary determined using the inconsistencies of the underlying image statistics as grayscale images. Rössler *et al.* [27] constructed a dataset for face manipulation detection, published it, and presented experimental results using existing detection techniques. They showed that extracting only the facial region and using it as an input image improves performance. The authors demonstrated that XceptionNet [14] exhibited the

highest performance among the detection networks used in their experiment.

Most existing methods detect specific types of deepfake or use only one strategy to detect deepfakes. However, because the deepfake types and generation methods are diverse, detecting fake faces is difficult using a single detection strategy. Furthermore, using a specific feature to detect deepfakes means that detection can be easily avoided. The proposed method combines various strategies into a single model for stable deepfake detection.

FIGURE 1. Sample dataset images.

TABLE 1. Descriptions of datasets.

deepfake Type	Method	Source	Train	Test
Face-swap	DeepFaceLab	YouTube	400 K	27 K
	-	DFDC	80 K	11 K
	-	Celeb-DF	82 K	5 K
Puppet-master	ICface	MegaFace	63 K	7 K
	Face2Face	FaceForensics	351 K	77 K
Attribute-change	Glow	CelebA-HQ	301 K	33 K
Fake total	-	-	1278 K	160 K
Original	-	YouTube	170 K	25 K
	-	DFDC	20 K	2 K
	-	Celeb-DF	82 K	5 K
	-	MegaFace	63 K	7 K
	-	FaceForensics	350 K	77 K
	-	CelebA-HQ	34 K	4 K
	-	VGG	558 K	40 K
Original total	-	-	1278 K	160 K

III. DATASETS

The datasets were collected or generated using representative methods, as shown in Table 1. The original and deepfake image datasets were preprocessed for face region detection and cropping. In dataset generation, deepfake images are created by applying various deepfake generation methods to the original images. Because deepfake images are generated using various combinations of original images, the number of deepfake images is relatively larger than that of the original images. Therefore we included the VGG face dataset [34] to match the numbers in training and testing equally. Figure 1 represents the original and fake sample images of each dataset.

A. GENERATING FACE-SWAP IMAGES

We generated face-swap images using a synthesis application that DeepFaceLab [7] provides. Because this method is based on an autoencoder with a GAN, numerous facial images of a specific person are required. The greater the variety and quantity of a specific person, the better the output quality. Thus, we collected 40 videos from five different people (Donald J. Trump, Moon Jae-in, Xi Jinping, Abe Shinzo, and Kim Jong-un). From these videos, we collected approximately 195 K cropped facial images and then trained the face-swap models for each person. After training, we generated 428 K face-swap images for all five people for swappable cases (${}_5P_2 = 20$). We also included 91 K face-swap images from the deepfake Detection Challenge (DFDC) [35] training dataset and 87 K face-swap images from Celeb-DF [6], which showed low distortions from various synthesis methods.

B. GENERATING PUPPET-MASTER IMAGES

To build puppet-master datasets, we used the ICFace [9] model. We used 3,521 frames extracted from randomly selected videos in the VoxCeleb2 dataset [36] as the target image datasets. For each image, attribute vectors representing the head poses and AUs of 17 facial muscles [37] were extracted using OpenFace [38]. We then randomly sampled 70,420 images from MegaFace [39] and generated 70,420 fake images using an attribute vector. We also included 428 K images from the FaceForensics [19] dataset, which was generated using Face2Face [20], a technique for facial reenactment manipulation.

C. GENERATING ATTRIBUTE-CHANGE IMAGES

We built attribute-change datasets using the Glow [11] model. We randomly selected 38 K images from CelebA-HQ [40] as source images. In total, 13 relatively valid and natural attributes (5_o_Clock_shadow, Bags Under Eyes, Bald, Black Hair, Blond Hair, Bushy Eyebrows, Chubby, Heavy Makeup, Male, No Beard, Rosy Cheeks, Smiling, and Young) were chosen for the transformation. Consequently, we obtained 334 K face images transformed from 38 K images.

IV. DEEPFAKE IMAGE DETECTION

A. TRACES OF DEEPFAKE

1) RESIDUAL NOISE

Once an image pixel is modified (or synthesized) by image operations, its relationship with its neighbor is expected to change, yielding traces with a periodicity that depends on image operations. To analyze traces of image generation, residual-domain approaches have been investigated for image forgery detection [41]. In deepfake detection, residual noise is a transfiguration trace generated by passing it through an autoencoder and GAN network filters [42].

Image residuals are affected by the transformation methods rather than the image content. Therefore, we propose a feature extraction method using a deep learning model that can focus on the residual noise of the deepfake operation. We propose the use of SRNet [43] to capture residual noise traces in deepfake images. SRNet, which is designed for steganalysis,

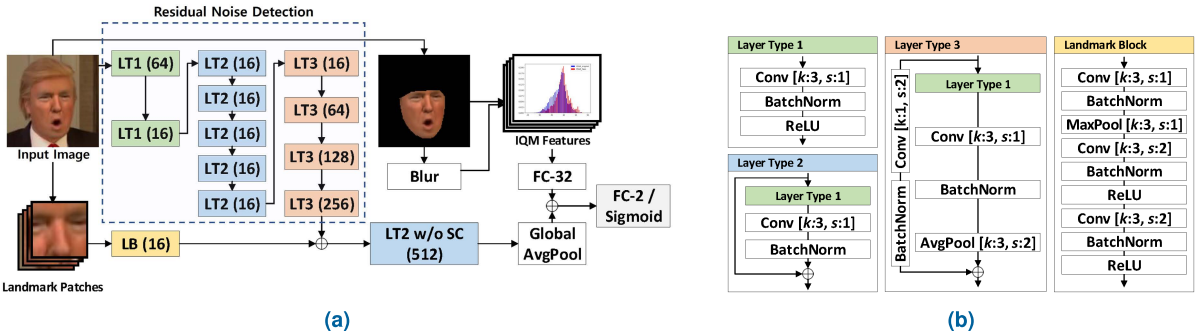

FIGURE 2. Proposed network architecture for deepfake detection. (a) Detection network structure, (b) details for layer types (LTs) and landmark block (LB). SC in (a) is an abbreviation for skip connection. In (a), the base network composed of Layer type 1-3 detects residual noise appearing throughout the image.

is used to concentrate on fine signals by excluding pooling layers at the front of the network. As the technique for deepfake generation has developed, distinguishing it from its content becomes more difficult. Differences between real and fake images, such as unnatural face collapse, edge of the pasted face, and unnatural eye and mouth behavior, will be minimal in high-level features. Therefore, detecting the residual noise caused by synthesis operations is an important factor in deepfake detection.

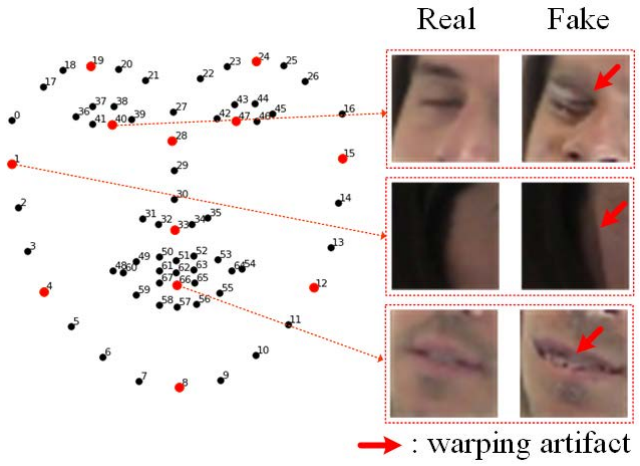

FIGURE 3. Landmark locations and patch examples. The red dots indicate the landmark locations used in our method.

2) WARPING ARTIFACTS

Owing to inaccurate predictions in facial geometry and light, deepfake images generate warping artifacts that are useful high-level features for detecting deepfake. For example, pupils and teeth, which require detailed expressions, are visually distorted in deepfake images. In addition, boundary artifacts typically appear on the forehead, chin, or edges of the face because of pasting. Warping artifacts appear because of the limitations of deepfake generation technology. The difference in the size and position of facial components, facial color, head angle, facial expression, and lighting conditions between the source and target can cause warping artifacts. Furthermore, a limited number of photos or videos and a lack of training can also cause warping artifacts. Warping artifacts tend to appear in semantic areas of the face, as shown in Fig. 3. Therefore, we extracted landmark image patches from semantic face regions to focus on warping artifacts.

3) BLUR EFFECTS

In [26], a deepfake-like dataset was created by blurring facial regions in images. From this, we inferred that a blur-like effect exists in deepfake images. This is because of the resolution inconsistency and postprocessing that occurs during the deepfake generation process. The resolution of the inputs and outputs of deepfake networks is typically fixed, whereas the resolution of the source or target image is not fixed. Furthermore, the size of the face in the image varied according to the distance from the camera. Therefore, face image resizing occurs frequently in the deepfake generation process, and causes interpolation and blur-like effects.

Owing to the limitations of deepfake generation techniques, the output of a deepfake typically has unnatural features. In particular, the output of a deepfake generation network often leaves the boundary owing to the discontinuity between the source and target faces [6]. Textured noise often occurs in the output. Therefore, postprocessing typically considers blurring for naturality, among other methods. Considering this, blur-like effects often occur in deepfake images, and we exploited this as a trace of deepfake. Blurring does not make a significant difference when applied to an already-blurred image. Based on this observation, we applied a Gaussian filter to face-only images and compared them to blurred face-only images using IQM tools.

As shown in Table 2, we used the following 17 IQM tools: Laplacian blur variance (LPV), high-low frequency index (HLFI) [48], spectral phase error (SPE) [49], spectral magnitude error (SME) [49], gradient-magnitude error (GME) [44], gradient phase error (GPE) [44], structural content (SC) [46], average difference (AD) [46], mean square error (MSE) [47], signal-to-noise ratio (SNR) [50] in dB, normalized absolute error (NAE) [46], peak signal-to-noise

TABLE 2. Descriptions of image quality measurements used for trace blur effects.

Name	Description	Name	Description
Laplacian blur Variance	$LPV(I) = Var(L(I))$	Gradient-Magnitude Error [44]	$GME(I,\hat{I}) = \frac{1}{nm}\sum^{n}\sum^{m}(\lvert G^{M}_{i,j} - \hat{G}^{M}_{i,j}\rvert)$
Normalized Cross-Correlation [45]	$NCC(I,\hat{I}) = \frac{\sum^{n}\sum^{m} I_{i,j}\hat{I}_{i,j}}{\sum^{n}\sum^{m} I^{2}_{i,j}}$	Gradient Phase Error [44]	$GPE(I,\hat{I}) = \frac{1}{nm}\sum^{n}\sum^{m}(\lvert G^{P}_{i,j} - \hat{G}^{P}_{i,j}\rvert)$
Structural Content [46]	$SC(I,\hat{I}) = \frac{\sum^{n}\sum^{m} I^{2}_{i,j}}{\sum^{n}\sum^{m} \hat{I}^{2}_{i,j}}$	Mean Square Error [47]	$MSE(I,\hat{I}) = \frac{1}{nm}\sum^{n}\sum^{m}(I_{i,j} - \hat{I}_{i,j})^{2}$
Normalized Absolute Error [46]	$NAE(I,\hat{I}) = \frac{\sum^{n}\sum^{m}\lvert I_{i,j} - \hat{I}_{i,j}\rvert}{\sum^{n}\sum^{m}\lvert I_{i,j}\rvert}$	Laplacian MSE [46]	$LMSE(I,\hat{I}) = \frac{\sum^{n}\sum^{m}(L(I_{i,j}) - L(\hat{I}_{i,j}))^{2}}{\sum^{n}\sum^{m} L(I_{i,j})^{2}}$
High-Low Frequency Index [48]	$HLFI(I) = \frac{\lVert F^{l}(I)\rvert - \lvert F^{h}(I)\rVert}{\sum_{u=0}^{n-1}\sum_{v=0}^{m-1}\lvert F_{u,v}(I)\rvert}$	Spectral Magnitude Error [49]	$SME(I,\hat{I}) = \frac{1}{nm}\sum^{n-1}\sum^{m-1}\lVert F_{u,v}(I)\rvert - \lvert F_{u,v}(\hat{I})\rVert^{2}$
Maximum Difference [46]	$MD(I,\hat{I}) = \max\lvert I_{i,j} - \hat{I}_{i,j}\rvert$	Spectral Phase Error [49]	$SPE(I,\hat{I}) = \frac{1}{nm}\sum^{n-1}\sum^{m-1}\lvert\varphi_{u,v}(I) - \varphi_{u,v}(\hat{I})\rvert^{2}$
Average Difference [46]	$AD(I,\hat{I}) = \frac{1}{nm}\sum^{n}\sum^{m}(I_{i,j} - \hat{I}_{i,j})$	Signal-to-Noise Ratio [50]	$SNR(I,\hat{I}) = 10\log_{10}\frac{\sum^{n}\sum^{m} I^{2}_{i,j}}{\sum^{n}\sum^{m}(I_{i,j} - \hat{I}_{i,j})^{2}}$
Peak Signal-to-Noise Ratio [51]	$PSNR(I,\hat{I}) = 10\log_{10}\frac{255^{2}}{MSE(I,\hat{I})}$	R-Averaged Max Difference [47]	$RAMD(I,\hat{I},R) = \frac{1}{R}\sum_{r=1}^{R}\max_{r}\lvert I_{i,j} - \hat{I}_{i,j}\rvert$

ratio (PSNR) [51], Laplacian MSE (LMSE) [46], maximum difference (MD) [46], R-averaged maximum difference (RAMD) [47], normalized cross-correlation (NCC) [45], and visual information fidelity (VIF) [52]. Owing to the complexity of the formula, VIF, which is an image quality assessment index that uses natural scene statistics to quantify the loss of image information, is excluded in Table 2. In Table 2, G is the gradient of an image, and G^M and G^P denote the magnitude and phase of G, respectively. Moreover, F indicates the Fourier transform operation, and F^l and F^h denote the low and high frequencies in the Fourier domain, respectively. In addition, φ indicates the phase in the Fourier domain, and L is the Laplacian filter such that $L(I_{i,j}) = I_{i+1,j} + I_{i-1,j} + I_{i,j+1} + I_{i,j-1} - 4I_{i,j}$.

Figure 4 is an example of the histograms for two IQM feature values: PSNR and LMSE. The red area represents the feature values of deepfake images, and the blue area represents the originals. In PSNR, a higher value represents a smaller difference between the two compared images. A higher LMSE value represents a larger difference. We found that deepfake images tend to have a blurred effect because they are less affected by the Gaussian filter. Therefore, we applied IQM features to the proposed network to capture blur-like traces in deepfake images.

B. DETECTION NETWORK

Figure 2 illustrates the overall architecture of the proposed network for deepfake detection, where LT and LB denote the layer type and landmark block, respectively. The numbers in parentheses are kernel numbers.

To capture residual noise, we propose adopting the SRNet architecture [43] as the base network. Primarily, SRNet is used in steganalysis to concentrate on fine signals at the pixel level. The key method is to not reduce the dimensions at the front of the network by excluding the pooling. In the experimental section, we demonstrate that SRNet can effectively capture noise traces in deepfake images. The base network is optimized to detect fine signals but does not focus on high-level features such as warping artifacts.

To detect warping artifacts, we extracted 14 landmark patches where the warping artifacts primarily appeared and used them as the input of the LB. Unlike the front part of the base network, a pooling layer was added to the LB because it does not need to preserve residual noise.

The output of the LB was concatenated with the output of LT 3 to deliver the warping artifact information. After LT 2, without a skip connection, we used global average pooling to prevent overfitting and reduce the number of neurons in the fully connected layer. Finally, to detect traces of blur effects, the 17 IQM features computed from the target and blurred images were concatenated with the output of the global average pooling after passing through a fully connected layer. Subsequently, the output passed through the last fully connected layer for two-class classification.

V. EXPERIMENTAL RESULTS

We used 2.88 M images to train and test the proposed network. The size of the fake and original training sets were 2.55 M. For the tests, 320 K fake and original images were used. The training and tests sets do not include the same subject such that training data leakage does not occur during the training process. For each deepfake dataset experiment, the original images were randomly imported such that the ratio of fake to original images was 1:1. We used the Dlib [53] library for preprocessing. Pre-processing included facial landmark detection and region cropping. As described in Fig. 3, 12 facial landmark patches were made around the landmark point of size 32×32.

The resolution of the input images was 128×128. While constructing the deepfake image dataset for our experiment, we collected deepfake images from FHD (1920×1080) and videos from the DFDC [35] and FaceForensics [19] dataset ($> 640 \times 480$). We then cropped the facial area from the images in these datasets, which reduced the input image size. Because we pre-extracted the facial area, the input size

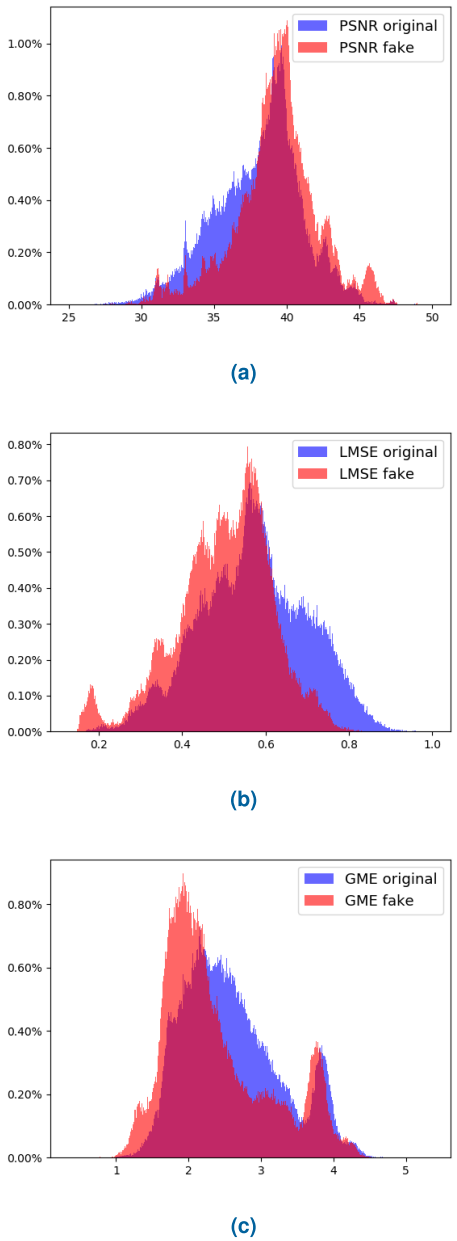

FIGURE 4. Histograms of IQM features for original and deepfake images: (a) PSNR, (b) LMSE, and (c) GME. The blurring operator causes a smaller effect on deepfake images than on non-fake images.

required to be considerably smaller than the original size. In addition, most of the cropped facial images had a size of 128 × 128. Therefore, we unified the image input size with this value.

The model was trained for 10 epochs, with a batch size of 32. The binary cross-entropy (BCE) loss function was used, with a learning rate of 0.001. We used the Adam optimizer was used for a smooth learning rate adjustment and initialized the weight using Kaiming initialization [54].

A. COMPARISON OF THE RESIDUAL-NOISE DETECTOR

Residual noise generated when producing deepfakes is fine noise at the pixel level. To detect this, we propose the use of SRNet, which was designed for steganalysis. To demonstrate the efficiency of the residual noise feature in deepfake detection, a comparative experiment was conducted with existing well-known convolutional neural network (CNN) models. For comparison, we trained and tested VGG [55], ResNet [56], DenseNet [57], XceptionNet [14], and SRNet [43] using our datasets. The environmental setting was the same as in the proposed network. Table 3 lists the accuracy of the CNN models for each deepfake type. SRNet, which focuses on fine noise, exhibits better performance than existing CNN models that analyze high-level features. The puppet-master technique does not change the face of the target person but changes only the expression of the target. Therefore, a significant difference in appearance was not observed compared with the other deepfake generation methods. This is because the puppet-master technique has the largest difference in detection performance between SRNet and other network models.

TABLE 3. Accuracy (%) of network models for each type of deepfake.

	FS.[i]	PM.[ii]	AC.[iii]	Total
VGG [55]	88.20	84.49	90.32	86.70
ResNet [56]	89.13	81.19	89.10	84.95
DenseNet [57]	88.19	79.24	90.61	84.00
Xception[14]	90.29	87.92	92.61	89.54
SRNet[43]	**95.51**	**94.32**	**95.64**	**94.96**

[i]FS. = Face-swap, [ii]PM. = Puppet-master, [iii]AC. = Attribute-change

TABLE 4. Boosted performance by adjusting landmark patches (LM) and image quality measurement (IQM).

	FS.[i]	PM.[ii]	AC.[iii]	Total
Xception	90.29	87.92	92.61	89.54
Xception + IQM	92.14	90.60	93.75	91.69
Xception + LM	91.44	90.55	93.67	91.42
Xception + LM + IQM	93.39	91.93	95.31	93.04
SRNet	95.51	94.32	95.64	94.96
SRNet + IQM	97.00	96.74	97.44	96.94
SRNet + LM	96.92	**96.87**	97.95	97.11
SRNet + LM + IQM	**97.83**	96.68	**98.23**	**97.33**

[i]FS. = Face-swap, [ii]PM. = Puppet-master, [iii]AC. = Attribute-change

B. BOOSTING PERFORMANCE USING THE PROPOSED FEATURES

We propose deepfake image detection using warping artifacts, blur effects, and residual noise. Landmark patches were applied to the network to detect warping artifacts and IQM features between images, and blurred images were used to capture the blur effect. For blurring, a Gaussian blur filter with a kernel size of 3, and a Gaussian kernel standard deviation in the X direction of 0.5 was used. We tested the effectiveness of our strategy by applying it to XceptionNet and SRNet, which exhibited high performance in previous experiments. In XceptionNet, we concatenated the result of the landmark block to the middle of the exit flow, and concatenated IQM features to the input of the final fully connected layer. Table 4

TABLE 5. Accuracy (%) and AUROC of detection models for each type of deepfake. (Accuracy/AUROC).

	Face-swap			Puppet-master		Attribute-change	Total
	DFL.[i]	DFDC	Celeb-DF	ICFace	FF.[ii]	Glow	
Li *et al.* [26]	86.15/.9621	71.64/.6717	63.69/.6913	63.86/.2121	75.36/.8282	44.36/.3619	64.22/.6753
Afchar *et al.* [25]	89.85/.9879	82.53/.8738	77.26/.8477	88.91/.9274	85.74/.9425	90.66/.9246	87.12/.9376
Rössler *et al.* [27]	90.36./9596	90.35/.9488	89.82/.9422	74.56/.8688	89.15/.9432	92.61/.9063	89.54/.9354
Ours1 (base+IQM)	98.75/.9999	93.30/.9915	**95.43/.9938**	97.70/.9971	96.64/.9871	97.44/.9939	96.94/.9916
Ours2 (base+LM)	**98.77/.9999**	95.32/.9945	90.33/.9863	97.81/.9968	**96.78/.9893**	97.95/.9946	97.11/.9927
Ours3 (base+LM+IQM)	98.59/.9999	**97.35/.9969**	94.78/.9925	**97.96/.9983**	96.56/.9922	**98.23/.9953**	**97.33/.9947**

[i]DFL. = DeepFaceLab, [ii]FF. = FaceForensics

shows the effect of applying warping artifact and blur-effect features to the network. The results show that each strategy is effective in detecting deepfakes. Furthermore, the network, which included the combination of each method, showed the highest performance on the total dataset. The total dataset refers to the dataset including face-swap, puppet-master, and attribute-change.

C. COMPARISON TEST AND DISCUSSION

We conducted a comparative experiment between the proposed method and other deepfake-detection techniques. In the experiment, SRNet was used as the base network for the proposed method. Li and Lyu [26], Afchar *et al.* [25], and Rössler *et al.* [27] with XceptionNet, which exhibited the highest performance on FaceForensic++, were used for the comparison. The proposed models, XceptionNet and Meso-4, were trained using our datasets. For ResoNet, we used a pretrained model for the tests.

Table 5 lists the accuracy and area under the receiver operating characteristics (AUROCs) of the network models for each type of deepfake, and Fig. 5 illustrates the receiver operating characteristic (ROC) curves for each network model per dataset. Our proposed networks have higher accuracies and AUROCs than existing deepfake-detection networks in all datasets. The SRNet+IQM method exhibited a higher performance than the SRNet+LM+IQM method on datasets that have relatively more blur-like traces, such as the Celeb-DF dataset. Similarly, the base+LM method was more effective than the SRNet+LM+IQM method on some datasets that have few traces of blurring and more traces of warping artifacts, such as the FaceForensics dataset. When the accuracy of SRNet+LM or SRNet+IQM is the highest, the gap between the highest accuracy value and the accuracy value of the SRNet+LM+IQM method is always small. However, on the DFDC and Celeb-DF datasets, the SRNet+LM+IQM method exhibited an accuracy approximately 4% higher than that on SRNet+LM and SRNet+IQM respectively. That is, the detection performance of the SRNet+LM and SRNet+IQM methods depends relatively more on the deepfake generation algorithm. However, the SRNet+LM+IQM method exhibits stable detection performance for various deepfake datasets, which is why the SRNet+LM+IQM method shows the highest accuracy and AUROC values for the entire dataset.

For time consumption, the IQM feature extraction of our method required 1.14 seconds, and 1.06 seconds to extract the landmark patches, which can be drastically shortened using parallel calculations. Table 6 lists the network execution times (based on NVIDIA RTX 2080Ti GPU) and number of network parameters. Compared to XceptionNet, which showed the best performance on FaceForensics++, our final network (SRNet+LM+IQM) showed better performance with approximately 60% of the execution time and 20% of the number of parameters. Furthermore, the additional number of network parameters and execution time required to detect blurry traces and warping artifacts were approximately 157 K and 0.001 s, respectively.

Limitations on the state-of-the-art deepfake detector have been reported, including a lack of generalization by focusing on specific artifact identification and rapid performance decreases as deepfake quality improves [58]. Addressing this, this study alleviated the generalization problem by detecting deepfakes using various strategies, and insignificant performance degradation was observed when experimenting with sophisticated deepfake images using the latest deepfake generation methods and datasets. Despite these strengths, some case predictions still fail. For example, an original image is determined a deepfake because of the rapid movement of a face or overlapped image that occurs when the scene changes when dividing a video into a frame-by-frame sequence. In addition, owing to the strategy of detecting blur effects, blurred images that are not deepfakes are sometimes incorrectly detected. Signal-processing attacks, such as compression-quality changes and noise additions, can also interfere with detection.

We discuss adversarial attack in terms of noise addition. Adversarial attacks are methods that cause errors in a model by adding perturbations to the image. This perturbation is a type of noise that is extremely small and has a fatal effect on the model [59]. Szegedy *et al.* [59] indicated that deep neural network (DNN)-based models are vulnerable to adversarial attacks, and [58] mentioned the problems and limitations of adversarial attacks when using deep learning models to detect deepfakes. Currently, studies on defending against adversarial attacks [60], [61] and attack defense techniques [62], [63] are being actively studied. In future studies, we can improve the robustness against adversarial attacks using methods such as adversarial training [59] and defensive distillation [61].

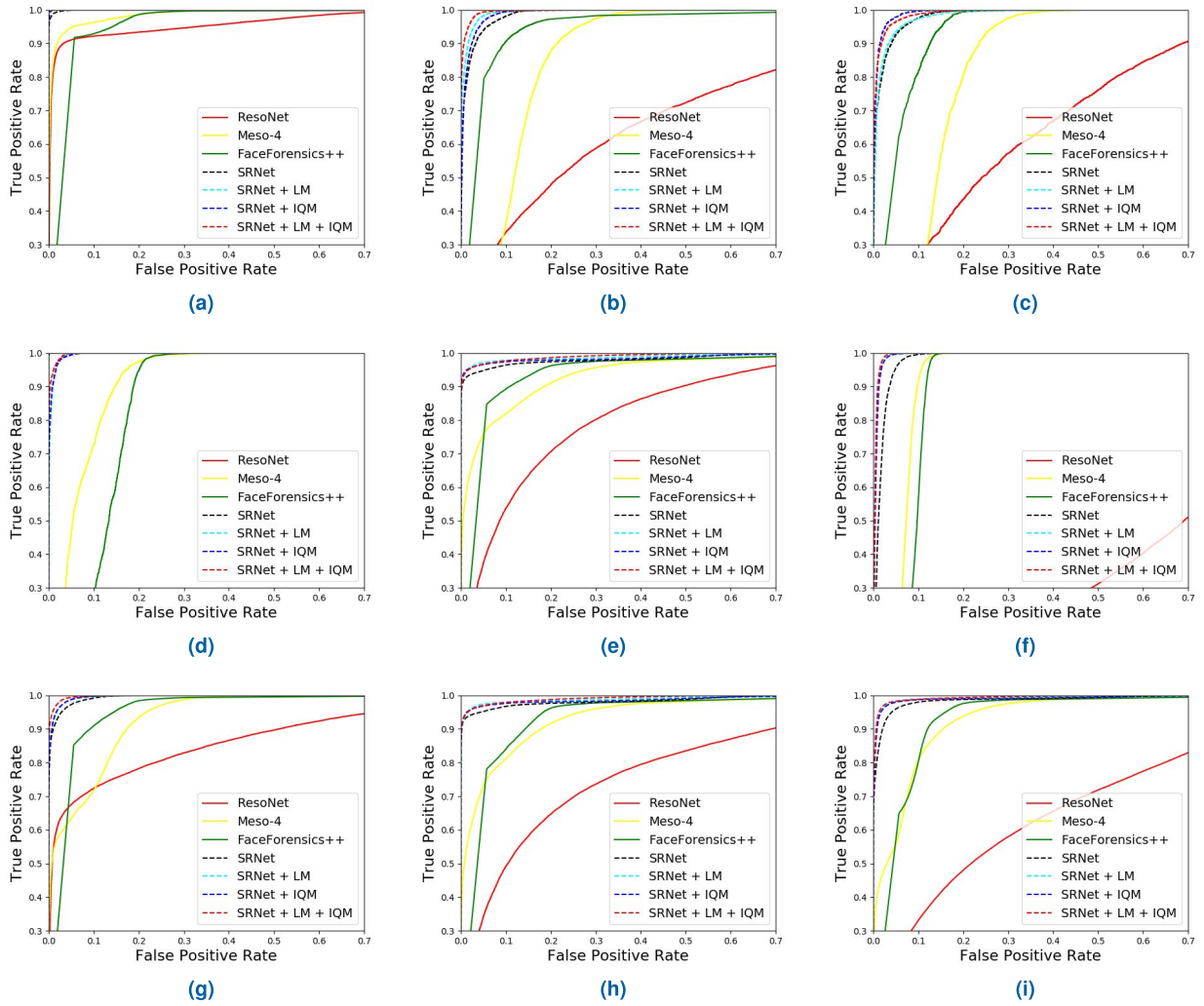

FIGURE 5. ROC plots for the network models of each deepfake type. (a)-(i) are graphs for the DeepFaceLab, DFDC, Celeb-DF, ICFace, FaceForensics, Glow, face-swap (DeepFaceLab, DFDC, Celeb-DF), puppet-master (ICFace, FaceForensics), and total datasets, respectively.

TABLE 6. Execution time and number of parameters of deepfake detection networks.

	ResoNet	Meso-4	FaceForensics++	SRNet	SRNet+LM	SRNet+IQM	SRNet+LM+IQM
Execution time (s)	0.1648	0.0015	0.0111	0.0057	0.0066	0.0061	0.0068
Num of parameters	23,512 K	15 K	20,811 K	4,779 K	4,935 K	4,780 K	4,936 K

Finally, methods that are more advanced than the *dlib* we used in the landmark patch extraction process are available. We can expect to further improve performance using state-of-the-art face detection technology, such as Mediapipe [64], in future studies.

VI. CONCLUSION

In this study, we proposed a generalized detection method to detect three types of deepfake techniques: face swap, puppet-master, and attribute change. We exploited three types of common traces (residual noise, warping artifacts, and blur effects) generated by the deepfake process. We applied them to the proposed network for deepfake detection. First, a network designed for steganalysis was adopted as the base network to detect residual noise. Second, landmark patches were extracted from the semantic facial region to detect warping artifacts, which are unnatural high-level features. Finally, we applied IQM features to capture the statistical characteristics of the blur-like effects of a deepfake. The results revealed that each detection strategy is effective, and the performance of the proposed network is superior to that of existing networks.

From an additional perspective, we focused on determining deepfakes' common traces, which are difficult to bypass. Because detecting image-based traces is more difficult to bypass than detecting traces of time-based inconsistencies in

deepfakes, we targeted image features. Because a deepfake video inherits residual features from image operations, our approach can be directly adopted for deepfake video detection pipelines based on frame-by-frame detection. Based on the proposed method, we plan to expand this study to include a deepfake video detection method. We hope this method is robust against signal- and time-based attacks.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[2] S. Wen, W. Liu, Y. Yang, T. Huang, and Z. Zeng, "Generating realistic videos from keyframes with concatenated GANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2337–2348, Aug. 2019.

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

[4] J. Damiani, "A voice deepfake was used to scam a CEO out of $243,000," *Forbes*, Sep. 2019.

[5] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.

[6] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3207–3216.

[7] *DeepFaceLab*. Accessed: Nov. 12, 2020. [Online]. Available: https://github.com/iperov/DeepFaceLab

[8] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.

[9] S. Tripathy, J. Kannala, and E. Rahtu, "ICface: Interpretable and controllable face reenactment using GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3385–3394.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[11] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10215–10224.

[12] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "DeepFake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[13] Y. Li, P. Sun, H. Qi, and S. Lyu, "Toward the creation and obstruction of deepfakes," in *Handbook of Digital Face Manipulation and Detection*. Cham, Switzerland: Springer, 2022, pp. 71–96.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[15] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3677–3685.

[16] R. Natsume, T. Yatagawa, and S. Morishima, "RSGAN: Face swapping and editing using face and hair representation in latent spaces," 2018, *arXiv:1804.03447*.

[17] L. Tran and X. Liu, "On learning 3D face morphable model from in-the-wild images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 157–171, Jan. 2019.

[18] Z. Geng, C. Cao, and S. Tulyakov, "3D guided fine-grained face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9821–9830.

[19] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," 2018, *arXiv:1803.09179*.

[20] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[21] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 818–833.

[22] J.-U. Hou and H.-K. Lee, "Detection of hue modification using photo response nonuniformity," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1826–1832, Aug. 2017.

[23] J. Wang, T. Li, X. Luo, Y.-Q. Shi, and S. K. Jha, "Identifying computer generated images based on quaternion central moments in color quaternion wavelet domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2775–2785, Sep. 2019.

[24] J. Wang, H. Wang, J. Li, X. Luo, Y.-Q. Shi, and S. K. Jha, "Detecting double JPEG compressed color images with the same quantization matrix in spherical coordinates," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2736–2749, Aug. 2020.

[25] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[26] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 1–7.

[27] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[28] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.

[29] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2019, pp. 83–92.

[30] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8261–8265.

[31] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," 2018, *arXiv:1808.07276*.

[32] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation," in *Proc. 20th Irish Mach. Vis. Image Process. Conf. (IMVIP)*, 2018, pp. 133–136.

[33] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5001–5010.

[34] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[35] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[36] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, Sep. 2018, pp. 1086–1090.

[37] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[38] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," CMU School Comput. Sci., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-16-118, 2016.

[39] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4873–4882.

[40] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[41] H. Li, W. Luo, X. Qiu, and J. Huang, "Identification of various image operations using residual-based features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 31–45, Jan. 2018.

[42] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7556–7566.

[43] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1181–1193, May 2019.

[44] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.
[45] J. Galbally and S. Marcel, "Face anti-spoofing based on general image quality assessment," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1173–1178.
[46] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
[47] I. Avcıbaaş, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *J. Electron. Imag.*, vol. 11, no. 2, pp. 206–223, Apr. 2002.
[48] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *Proc. Int. Workshop Quality Multimedia Exp.*, Jul. 2009, pp. 64–69.
[49] N. B. Nill and B. Bouzas, "Objective image quality measure derived from digital image power spectra," *Proc. SPIE*, vol. 31, no. 4, pp. 813–826, 1992.
[50] S. Yao, W. Lin, E. Ong, and Z. Lu, "Contrast signal-to-noise ratio for image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2005, p. 397.
[51] 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.
[52] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[53] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[57] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[58] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, Jan. 2022.
[59] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.
[60] R. E. Sutanto and S. Lee, "Real-time adversarial attack detection with deep image prior initialized as a high-level representation based blurring network," *Electronics*, vol. 10, no. 1, p. 52, Dec. 2020.
[61] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.
[62] G. Ryu, H. Park, and D. Choi, "Adversarial attacks by attaching noise markers on the face against deep face recognition," *J. Inf. Secur. Appl.*, vol. 60, Aug. 2021, Art. no. 102874.
[63] C. Bisogni, L. Cascone, J.-L. Dugelay, and C. Pero, "Adversarial attacks through architectures and spectra in face recognition," *Pattern Recognit. Lett.*, vol. 147, pp. 55–62, Jul. 2021.
[64] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Guang Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.

JIHYEON KANG received the B.S. degree from the School of Computer Science and Electrical Engineering, Handong Global University, South Korea, in 2015, and the M.S. and Ph.D. degrees from the Graduate School of Information Security, Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2017 and 2021, respectively. He is currently working at Webtoon AI, NAVER WEBTOON Corporation, South Korea. His research interests include machine learning and computer vision.

SANG-KEUN JI received the B.S. degree from the Department of Computer and Software Engineering, Kumoh National Institute of Technology, South Korea, in 2013, and the M.S. and Ph.D. degrees from the Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2015 and 2020, respectively. His current research interests include multimedia security and image processing.

SANGYEONG LEE is currently a Senior majoring in big data at Hallym University, South Korea. She is also an Undergraduate Researcher at the Multimedia Computing Laboratory. Her research interests include multimedia forensics, computer vision, and deep learning.

DAEHEE JANG received the Ph.D. degree in information security from KAIST, in 2019. He worked as a Postdoctoral Researcher at Georgia Tech, until 2020. He is currently an Assistant Professor with the Department of Security Engineering, Sungshin Women's University. He participated in various global hacking competitions (such as DEFCON CTF) and won several awards. He received the Special Prize from 2016 KISA Annual Event for finding 0-day security vulnerabilities in many software products. He is also the Founder of pwnable.kr wargame—an education platform for training hacking skills.

JONG-UK HOU received the B.S. degree in information and computer engineering from Ajou University, South Korea, in 2012, and the M.S. and Ph.D. degrees from KAIST, South Korea, in 2014, and 2018, respectively. He has been an Assistant Professor with the School of Software, Hallym University, since 2019, and a Principal Investigator of the Multimedia Computing Laboratory. His major interests include various aspects of information hiding, point cloud processing, computer vision, machine learning, and multimedia signal processing.