ForgeryNIR: Deep Face Forgery and Detection in Near-Infrared Scenario

Yukai Wang, Chunlei Peng, *Member, IEEE*, Decheng Liu, Nannan Wang, *Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— Deep face forgery and detection is an emerging topic due to the development of GANs. Face forgery detection relies greatly on existing databases for evaluation and adequate training examples for data-hungry machine learning algorithms. However, considering the wide application of face recognition in near-infrared scenarios, there is no publicly available face forgery database that includes near-infrared modality currently. In this paper, we present an attempt at constructing a large-scale dataset for face forgery detection in the near-infrared modality and propose a new forgery detection method based on knowledge distillation named cross-modality knowledge distillation aiming to use a teacher model which is pre-trained on the visible light-based (VIS) big data to guide the student model with a small amount of near-infrared (NIR) data. The proposed near-infrared face forgery dataset, named ForgeryNIR, contains a total of over 50,000 real and fake identities. A number of perturbations are applied to help simulate real-world scenarios. All source images in ForgeryNIR are collected from CASIA NIR-VIS 2.0, and fake images are generated via multiple GAN techniques. The proposed dataset fills the gap of face forgery detection research in the near-infrared modality. A comprehensive study on six representative detection baselines is conducted to evaluate the performance of face forgery detection algorithms in the NIR domain. We further construct a hard testing set, named ForgeryNIR+, which contains forged images that have bypassed existing face forgery detection methods. The proposed datasets will be publicly available and aim to help boost further research on face forgery detection, as well as NIR face detection and recognition.**

***Index Terms*— Near-infrared face, face forgery detection, deepfake.**

I. Introduction

RECENTLY, the rapid development of deep image generation technologies such as Generative Adversarial Networks (GANs) has aroused widespread public attention and interest. GAN was first proposed by Goodfellow [1] in 2014. After that, many researchers have also proposed many variants to improve the quality of the generated images, such as [2]–[5]. These efforts make face editing easier and relieve the users from the tedious manual editing process, thus greatly reduce the barriers for editors. However, such enabling technology can be used not only for mass entertainment but also for fake face manipulation. Face forgery is a technology that uses artificial intelligence technologies such as deep learning to generate fake visual identity information. General forgery methods include face swap, face attribute manipulation, face expression reenactment, and GAN-based fake generation. The abuse of face forgery technology on the Internet has led to a flood of fake videos and brings uncontrollable effects to society [6], such as damaging the reputation of celebrities,[1] affecting the image of politicians,[2] falsifying evidence, spreading rumors,[3] and interfering in elections,[4] *etc.* More specifically, in the face recognition scenario, [7] showed that the state-of-the-art face recognition systems based on VGG and Facenet neural networks are vulnerable to Deepfake videos. To deal with this situation, a lot of research efforts have been dedicated to the area of face forgery detection. Many researchers have proposed a series of countermeasures as well as provided high-quality datasets to evaluate the performance of face forgery detection algorithms and then help to improve the accuracy of these methods. For example, UADFV [8] contains 49 real videos and 49 fake videos which are generated by FakeAPP [9], but the forgery method is relatively simple. FaceForensics++ [10] expands the existing dataset in terms of scale and diversity, and applies four forgery technologies in it. Celeb-DF [11] is mainly proposed to solve the drawbacks of UADFV [8], FaceForensics++ [10], and other datasets, such as low image resolution, rough tampering traces, and excessive flickering of video faces. DeeperForensics-1.0 [12] is considered to be a large-scale face forgery dataset, containing 60,000 videos totaling 17.6 million frames. Furthermore, in DeeperForensics-1.0 [12], random perturbations are added in fake videos and the perturbed videos are counted as new fake videos. More recent datasets, such as ForgeryNet [13], FFIW-10K [14], DF-W [15], KoDF [16], adopt more advanced face forgery technology and can be applied to more forgery detection tasks. A lot of research efforts have been dedicated

Manuscript received July 28, 2021; revised November 10, 2021 and January 9, 2022; accepted January 19, 2022. Date of publication January 26, 2022; date of current version February 8, 2022. This work was supported in part by the Guangxi Natural Science Foundation Program under Grant 2021GXNSFDA075011; in part by the Key Research and Development Program of Shaanxi under Grant 2020ZDLGY08-08; in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; and in part by the National Natural Science Foundation of China under Grant 61922066, Grant 61876142, Grant 62036007, and Grant 62072356. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Domingo Mery. *(Corresponding author: Chunlei Peng.)*

Yukai Wang, Chunlei Peng, and Decheng Liu are with the State Key Laboratory of Integrated Services Networks, School of Cyber Engineering, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: ykwang.xidian@gmail.com; clpeng@xidian.edu.cn; dchliu@xidian.edu.cn).

Nannan Wang is with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: nnwang@xidian.edu.cn).

Xinbo Gao is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaoxb@cqupt.edu.cn).

Digital Object Identifier 10.1109/TIFS.2022.3146766

[1] https://www.bbc.com/news/technology-42912529

[2] https://www.bbc.com/news/av/technology-40598465

[3] https://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy

[4] https://edition.cnn.com/2019/06/22/politics/russia-fake-rubio-tweet/index.html

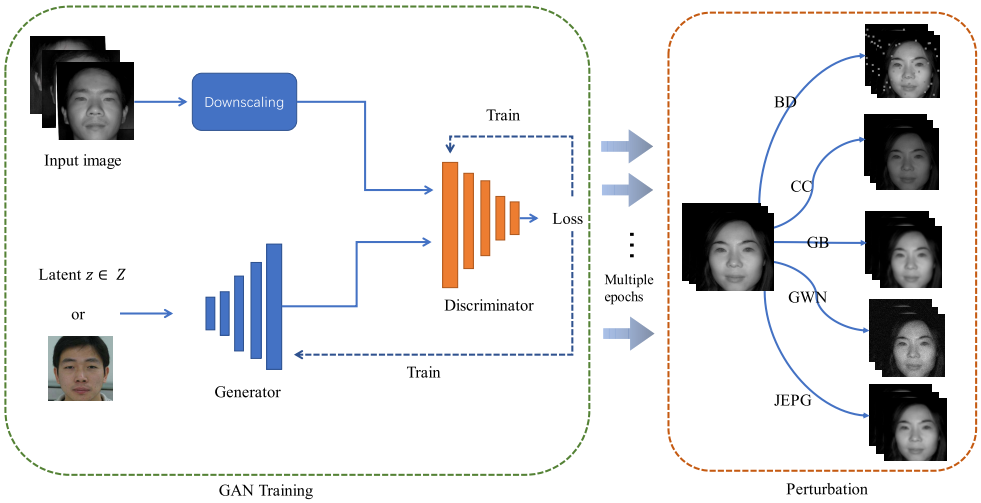


Fig. 1. Illustration of the proposed ForgeryNIR dataset for face forgery and detection in the near-infrared modality. BD, CC, GB, GWN, JPEG denote for the five types of perturbations in Table II.

to the field of face forgery detection. Even so, all existing datasets are confined to the visible light (VIS) modality. The large discrepancies between the near-infrared and visible light spectrum led to the poor forgery detection capabilities of the existing VIS forgery detection models in NIR scenarios. Furthermore, the NIR-based face recognition systems are also threatened by face forgery attacks. A lot of research efforts have been dedicated to presentation attacks and adversarial attacks, and most of them are based on research conducted within the modality of visible light (VIS). Even if some have also made some explorations in the near-infrared (NIR) modality, such as [17], [18]. However, the main contribution of these two works lies in adversarial samples in the near-infrared modality, rather than forgery samples. There is no exploration of GAN-based forgery faces under the NIR modality. However, near-infrared technology is widely used in face unlocking applications. For example, the research of "Near-infrared v.s. Visualization" aims to provide an alternative method to solve the illumination problem of the VIS face recognition system. [19] pointed out that due to the difference of spectral components between NIR and VIS modalities, visible light and near-infrared (VIS-NIR) face recognition is still a challenging task, and proposed a method of converting visible light into near-infrared modality to help the near-infrared face recognition system. The CASIA NIR-VIS 2.0 Face Database is proposed for supporting NIR-VIS face recognition. IPhone X uses Face ID[5] face unlocking function and uses a near-infrared camera to solve the lighting problem when unlocking. At the same time, near-infrared technology is also widely used in iris unlocking.[6] The iris is a ring-shaped film located between the black pupil and the white sclera on the surface of the human eye. Under the modality of infrared light, it can present rich visual features such as spots, stripes, and filaments, Coronal, crypt, *etc*. Therefore, whether it is a human face, an iris, or other identity information, it is likely to be forged in the near-infrared modality in the future. *Considering the wide application of authentication under the near-infrared (NIR) scenario, the latent threat of face forgery and detection in the NIR modality should be considered.* For example, a NIR-based access control system may encounter data poisoning attacks on facial information during system registration. In the authentication process, the NIR-based face unlocking technology used in the Xiaomi 8 phone was successfully cracked by simply using a black and white printer printout and adding shadows around the eyes and nose of the face to give it a more three-dimensional appearance. In addition, near-infrared sensor is often used for night video surveillance because it is robust to illumination. However, if the NIR surveillance images collected in the night scene are forged to create rumors out of nothing and instigate the spread of public opinion, it will damage the reputation of citizens [20]. The research on the authenticity of the NIR face is helpful to evaluate the fidelity of the generated NIR face, thereby providing augmented images for the training of the depth model in the heterogeneous face recognition research [21]. Our proposed forgery aims to address the security concerns of the aforementioned application scenarios in the NIR modality.

Face forgery detection relies greatly on existing databases for evaluation and adequate training examples for data-hungry machine learning algorithms. In order to solve the limitation of the current datasets in terms of modality diversity, we construct the first large-scale dataset in near-infrared (NIR) modality, named ForgeryNIR, to promote the empirical study of face forgery detection in NIR scenario. The ForgeryNIR dataset contains a total of over 50,000 real and fake identities for face forgery detection in the NIR modality. The main construction process of our dataset is shown in Fig. 1. We further construct a hard testing set, named ForgeryNIR+, which contains forged images that have bypassed existing face forgery detection methods. Compared with existing datasets, our ForgeryNIR has the following three advantages:

1) NIR modality. Because the face recognition system under the NIR scenario is usually robust to the lighting

[5]https://www.pocket-lint.com/phones/news/apple/142207-what-is-apple-face-id-and-how-does-it-work

[6]https://www.csoonline.com/article/3197933/hackers-easily-trick-iris-scanner-to-unlock-samsung-galaxy-s8.html

effect, it is widely applied in the dark situation and the environment where the lighting is varying. However, there is no face forgery and detection dataset in the NIR scenario. Considering the data poisoning threat to NIR based access control system during facial information registration, it is essential to evaluate face forgery and detection in NIR scenario. Therefore, our proposed dataset fills the gap of face forgery detection in the near-infrared modality.

2) High quality. We systematically conduct image quality assessment experiments on the fake images generated by various GAN methods ([2]–[5]) used in ForgeryNIR, and we use five image quality evaluation metrics, including two traditional non-reference image quality assessment algorithms [22], [23] and three types of deep learning-based models in NIMA [24] for neural image quality evaluation. It can be seen from the experimental results that the fake and the real near-infrared images contained in the ForgeryNIR have approximate quality values (shown in III-B). This means that regardless of human observation or distribution, the forged images in our dataset have high quality.
3) Rich diversity. We carefully apply five types of perturbations to manipulate the fake images at five intensity levels, leading to 25 perturbations in total to mimic real-world image processing and transmission situations. Furthermore, we propose a "multiple epochs" strategy to increase the diversity of the ForgeryNIR. Considering the intricacies of real-world scenarios, the forgery method selected by the forger and the number of training iterations of the model is uncontrolled. Using the dataset previously used for face forgery detection, where each forged image is generated using a certain fixed model may lead to overfitting of the trained model. In addition, in the pandemic time, we add a new perturbation like partial-face (due to face-mask) and make several experiments based on these faces wearing the mask. Furthermore, the images in the real-world scenario may be affected by more than one type of distortion. To better simulate the real-world condition, mixed types of distortion are applied to the forged faces. The details about the amount and distribution of the distortions applied in our ForgeryNIR dataset are presented in section II.

In summary, our contributions are as follows.

1) We present an attempt on constructing a large-scale dataset, named ForgeryNIR,[7] for face forgery detection in NIR modality.
2) The forged faces in our dataset are generated under NIR modality, with high quality and rich diversity, while a harder ForgeryNIR+ testing set is built containing fake faces that have bypassed existing face forgery detection models.
3) Extensive experiments are conducted to benchmark multiple deep learning-based face forgery algorithms on the proposed dataset, in order to help promote the research of near-infrared fake face detection in the future.
4) We propose a new forgery detection method base on knowledge distillation. The proposed cross-modality knowledge distillation (CMKD) method can achieve satisfactory performance on our novel topic of face forgery detection in NIR.

The remainder of this paper is organized as follows. Section II mainly introduces the current database and detection methods of face forgery detection. Section III introduces the data collection and generation process of our proposed database ForgeryNIR. At the same time, it also proposes the hard testing set ForgeryNIR+, and introduces the statistical characteristics of the dataset. Finally, we propose a new forgery detection method based on knowledge distillation, named cross-modality knowledge distillation. Section IV introduces the quantitative experiments on multiple forgery detection baselines and the qualitative experiments done by user study. Section V summarizes this paper and presents the outlook for the future.

II. Related Work

A. Summary of the Existing Face Forgery Detection Datasets

With the development of deep face forgery technology, deepfakes [25] have brought serious security risks to personal privacy and national security. In order to meet the challenge, researchers have done a series of studies on these manipulated videos and images. In addition, more and more high-quality datasets have been released to promote research in face forgery detection. Therefore, more and more researchers are eager for a large-scale dataset to further improve the generalization ability and robustness of the model. The current representative deep forgery detection datasets are shown in Table I.

Recently, many datasets have been proposed for face forgery detection. UADFV [8] contains a total of 98 videos with a ratio of 1:1 in which the real videos are collected from YouTube and the fake ones generated by FakeAPP [9]. Celeb-DF [11] is constructed to solve some of the drawbacks of the prior datasets, such as low image resolution, rough tampering traces, and excessive flickering of video faces. It contains 408 real videos which are collected from YouTube and 795 fake videos generated by a refined synthesis forgery algorithm. FaceForensics++ [10] is the first large-scale face forgery detection dataset, which contains 1,000 real videos, and uses four methods (DeepFakes [25], Face2Face [26], FaceSwap [27], Neural Texture [28]) to manipulate 1,000 fake videos with a total of 4,000 fake videos. DFDC [7] is a dataset provided by the Deep Forgery Detection Challenge Competition on Kaggle which containing 1,131 real videos and 4,113 fake videos and adding three types of perturbations to the dataset. DeeperForensics-1.0 [12] is a larger face forgery dataset, containing 60,000 videos with 17.6 million frames, where the ratio of real and fake videos is 1:5. The fake videos in DeeperForensics-1.0 are generated by a new end-to-end face-swapping method (DF-VAE). Considering the unpredictability and complexity under real-world forgery scenarios, the latent threat of multiple forged identities in one frame of the video should be considered. FFIW-10K [14] contains 10,000 real and 10,000 fake videos and is unique in its

[7] https://github.com/AEP-WYK/forgerynir

TABLE I
FACE FORGERY DETECTION DATASETS

Datasets	Numbers	Ratio (real: fake)	Multiple epochs	Numbers of perturbations	Modality
UADFV (2018) [8]	98	1:1	✗	0	VIS
DeepFake-TIMIT (2018) [31]	640	only fake	✗	0	VIS
Celeb-DF (2019) [11]	1,203	1:1.95	✗	0	VIS
FaceForensics++ (2019) [10]	5,000	1:4	✗	2	VIS
DFDC Preview Dataset (2019) [7]	5,214	1:3.6	✗	3	VIS
DeeperForensics-1.0 (2020) [12]	60,000	1:5	✗	35	VIS
KoDF (2021) [16]	237,942	1:2.8	✗	0	VIS
ForgeryNet (2021) [13]	221,247	1:1.2	✗	36	VIS
FFIW-10K (2021) [14]	20,000	1:1	✗	0	VIS
DF-W (2021) [15]	1,869	Only fake	✗	0	VIS
Ours	**50,000**	**1:4.1**	✓	**25**	**NIR**

real-world complexity. It is the first large-scale face forgery detection dataset for completely unconstrained multi-person facial forgery detection. The dataset generates each video with three face-swapping algorithms, including two learning-based methods (DeepFaceLab [29] and FSGAN [30]), and one graphic-based method (FaceSwap [27]). The videos in the dataset are longer comparing existing forgery datasets that relatively increase the difficulty of forgery detection when only key frames are forged. ForgeryNet [13] contains 99,630 real videos and 121,617 fake videos. The biggest advantage of this dataset is rich in annotations under both image and video-level, facilitating the application to more tasks (*i.e.*, image forgery classification, spatial forgery localization, video forgery classification, temporal forgery localization). KoDF [16] contains 175,776 fake videos and 62,166 real videos of 403 Korean characters, and it is proposed to balance the problem of fewer Asian faces in existing datasets.

These datasets indeed promote the development of the face forgery detection field. Considering the lack of near-infrared modality in existing datasets, we present an attempt at constructing a large-scale dataset for face forgery detection in NIR modality. The presented dataset, named ForgeryNIR, contains a total of over 50,000 real and fake identities. Several perturbations are applied to help simulate the real-world situation. Our ForgeryNIR dataset has three advantages: (1) NIR modality. (2) Rich diversity. (3) High quality, which are not considered in existing datasets.

B. Face Forgery Detection Methods

A carefully forged image or video may affect people's privacy, even endanger national security, and brings uncontrollable effects to society. Recently, many researchers have conducted a series of research on the problem of face forgery. Most of these methods regard the face forgery detection task as a common classification task, and the goal is to train a detection classifier with high accuracy and strong robustness.

In the early years, researchers have been focusing on graphics and traditional classification methods of and machine learning. [32] proposed a method that statistical features such as specular reflection, image quality distortion and color diversity can be combined and directly sent to SVM for binary classification. [33] pointed out that forged images are difficult to be distinguished in the RGB space, but the textures in other color spaces can be obviously different and presented a forgery detection method based on multi-level LBP features of human faces in HSV space and LPQ features in YCbCr space. Fernandes *et al.* [34] used heart rate biosignals to distinguish fake videos, and extracted the heart rate through three methods: changes in facial skin color caused by blood flow, average optical density of the forehead, and changes in Euler images. It is mainly based on the difference in heart rate distribution between normal videos and abnormal videos. Taking into account the difference in optical flow field caused by motion, [35] proposed a method of using PhotoPlethysmoGraphy (PPG) as a biosignal feature in blood pressure detection methods. This method extracts PPG signal features from multiple local areas of the face for fake face analysis. The latest research results in the field of computer vision show that convolutional neural networks (CNN) can effectively extract powerful and robust visual features and is superior in classification tasks when a significant amount of training samples is available. Therefore, forgery detection methods based on deep learning have attracted more and more attention. [36] provided a comprehensive review of face forgery and detection technology, and highlighted the improvement and challenges of the latest deepfake technology in forgery detection. [37] described a new strategy to remove GAN "fingerprints" from synthetic fake images based on autoencoders to deceive the facial manipulation detection system while maintaining the visual quality of the resulting images. [38] studied the detection performance of several image forgery detectors on forged face images, whether it is under ideal conditions or in the presence of compression which is routinely executed when uploading on social networks. [39] pointed out that the fake face images generated based on the GAN model often contain special fingerprint features and can be used for the detection of fake images. Inspired by this, the team of professor Larry Davis of the University of Maryland proposed a method to detect GAN forged face images based on GAN fingerprint features [40]. CNNDetection [41] found that even a classifier trained on images generated by one type of CNN model can show amazing generalization capabilities across datasets, network architectures, and training tasks. [42] found that GAN fake images often have obvious grid-like

Fig. 2. Example images selected from ForgeryNIR, where real images are collected in (a) CASIA NIR-VIS 2.0 and forgery images are generated by four types of GANs: (b) CycleGAN, (c) ProGAN, (d) StyleGAN, (e) StyleGAN2.

fake traces in the frequency domain. In terms of multi-target face forgery detection, [14] proposes a discriminative attention model for the classification and localization of face forgery.

Although the existing face forgery detection methods based on deep-learning have made remarkable progress, it still encounters some problems under near-infrared modality. With the wide application of face recognition under the NIR scenario, these algorithms only focus on visible light scenes and there is no publicly available face forgery database that includes the NIR modality faces. It is urgent to build a new dataset to deal with the lack of forgery faces in the NIR modality for model training. **Considering the realistic application value and research significance, we believe that it is the right time to call for more attention and research efforts to the area of face forgery detection under NIR modality.** Our ForgeryNIR enables us to explore this new research direction.

III. ForgeryNIR Dataset

In this section, we introduce our novel dataset based on the NIR modality. Fig. 2 shows example images selected from our proposed ForgeryNIR database. First, section III-A and section III-B introduces the data collection and forgery process respectively. Next, in order to better simulate the uncertainty and complexity of the forgery scene in reality, section III-C proposes the hard testing set, named ForgeryNIR+. At last, section III-D gives specific statistical indicators of ForgeryNIR and ForgeryNIR+. III-E proposes a novel method called cross-modality knowledge distillation (CMKD) which can achieve satisfactory performance on our novel topic of face forgery detection in NIR modality.

A. Collection of the Source Data

In practice, with the popularity of various sensors, near-infrared technology is widely used in many applications, especially in face recognition. Non-identity factors generally include illumination, posture, expression, accessories, *etc.*

TABLE II
Five Types of Perturbations in ForgeryNIR

Method	Perturbation Type
BD	Local block-based distortion
CC	Change in color contrast
GB	Gaussian blur
GWN	Gaussian white noise
JPEG	JPEG compression

In order to eliminate the influence of these non-identity factors, heterogeneous face biometrics which aims to match a pair of faces captured from two different modalities has attracted more attention in the face recognition community. Traditional face recognition systems use ordinary visible light face image recognition technology but are usually not robust to changes in ambient light. In scenes with large changes in ambient light, these methods can not perform well. Although the illumination preprocessing algorithm can eliminate the influence of illumination to a certain extent, NIR face recognition is a solution proposed to solve the illumination problem in face recognition. CASIA NIR-VIS 2.0 [43] is a typical dataset proposed for NIR-VIS recognition including 725 identities with 17,580 face images. and the ages of the identities are very wide, from children to old people. The faces in the NIR modality are selected as the training source data for forged face generation as well as the real NIR images in the proposed ForgeryNIR dataset.

B. Forged Face Generation

As there are currently no deepfake datasets under the NIR modality, and few people have done research on the forgery faces in the NIR modality. we refer to several of the latest datasets under VIS modality, such as FaceForensics++ (2019) [10], DeeperForensics-1.0 (2020) [12]. The fake faces in FaceForensics++ is generated using four methods based on face swap or face manipulation methods (*i.e.*, DeepFakes [25], Face2Face [26], FaceSwap [27], Neural Texture [28]). The fake videos in DeeperForensics-1.0 are generated by a new end-to-end face-swapping method (DF-VAE) with high robustness and scalability. Following the practices of the previous papers, we also adopt the generated strategy. For the generation of forgery faces, we train four GAN models *i.e.*, CycleGAN [2], ProGAN [3], StyleGAN [4], StyleGAN2 [5] with better current effects. CycleGAN is widely used for style transfer, so we use CycleGAN to generate fake face images here. A VIS face in CASIA NIR-VIS 2.0 [43] is selected as an input image, and the VIS modality of the source image is converted into a NIR mode through the CycleGAN model. The other three GANs are currently commonly used models for generating high-resolution images. The input of these models is random noise and the output is a fake image. ProGAN first manually reduces the training image to a very small initial resolution (only 4×4 pixels). It then create a generation and discrimination network with only a few network layers to synthesize the low-resolution image. Since the network is very small, its training is relatively faster, at this time the large-scale

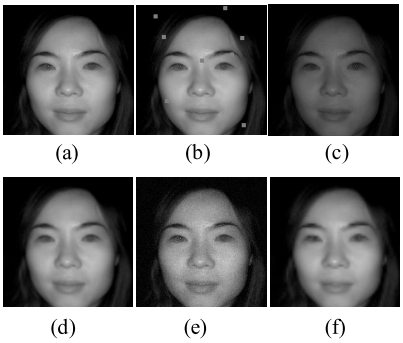


Fig. 3. Example images selected from ForgeryNIR dataset with different types of perturbations. (a) Fake NIR image without perturbations. (b) Local block-based distortion. (c) Change in color contrast. (d) Gaussian blur. (e) Gaussian white noise. (f) JPEG compression.

structure of the highly blurred image is learned. The ProGAN networks continuously increase the number of layers of the generation and discrimination network, and train the network until the resolution as we expected is 256×256 (here we use the same image resolution as in CASIA NIR-VIS 2.0 [43]). StyleGAN [4] proposes a new generator architecture that can control the high-level attributes of generated images, such as hairstyle, freckles, *etc*. StyleGAN2 [5] is proposed because of the flaws of StyleGAN that a small number of generated pictures have obvious water droplets and this drawback also exists on the feature map. By modifying the StyleGAN network structure and deleting the AdaIN operation, the problem is solved under the premise of ensuring the quality of generation. In this paper, we used the four GAN models mentioned above to generate four types of fake face images, which is defined as the subset **ForgeryNIR-std**.

With the improvement of current coding technology, although digital image transmission has become reliable, it is impossible to guarantee 100% distortion-free transmission. Almost everyone has encountered distortion caused by compression and data transmission errors. Many people experience distortions caused by image acquisition conditions-noise, blur and chromatic aberration. There are also distortions due to the specific operations of image processing such as denoising, averaging, and contrast changes. These types of perturbations are currently the mainstream strategies used in forgery datasets, such as FaceForensics++ (2019) [10], Deeperforensics-1.0 (2020) [12], and ForgeryNet (2021) [13] and these are also the perturbations that are prone to appear in real life. Therefore, a subset named **ForgeryNIR-rand** is constructed in our dataset by adding 5 common perturbations to the images to better imitate distorted scenes in the real world. Our dataset contains real and fake faces, following the principle of fairness, we add the same perturbation to them. Examples of adding five types of perturbations are shown in Fig. 3. The five types of perturbations are as follows:

1) Local block-based distortion: the distortion has been modeled in such a way that blocks of size 32×32 pixels that have arbitrary random color are placed in an image randomly or mainly in places where there is important information.
2) Change in color contrast: color contrast refers to the measurement of different brightness levels between the brightest white and the darkest black in the bright and dark areas of an image.
3) Gaussian blur: the distortion is also called Gaussian smoothing. It is usually used to reduce image noise and reduce the level of detail and also used in the preprocessing stage of computer vision algorithms to enhance the image effect of images at different scales.
4) Gaussian white noise: amplitude distribution obeys Gaussian distribution, and power spectral density obeys uniform distribution.
5) JPEG compression: JPEG was designed to compress color or gray-scale continuous-tone images of real-world subjects: photographs, video stills, or any complex graphics that resemble natural subjects.

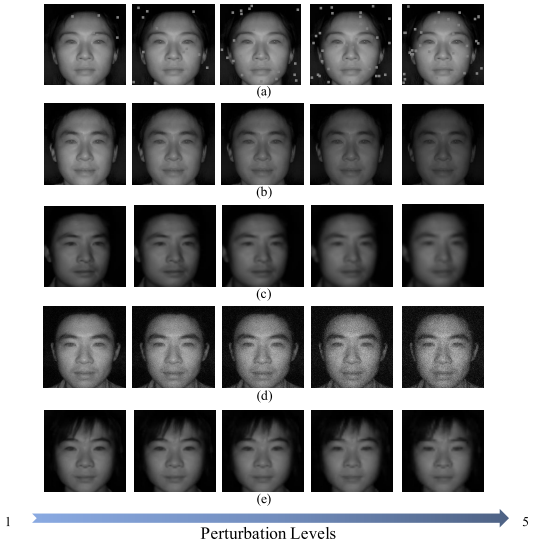


Fig. 4. Examples of visual quality after adding five types of perturbations: (a) Local block-based distortion. (b) Change in color contrast. (c) Gaussian blur. (d) Gaussian white noise. (e) JPEG compression.

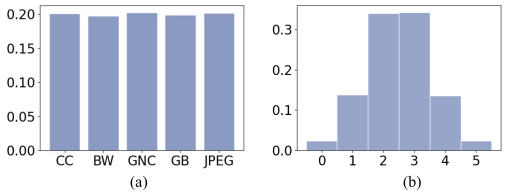


Fig. 5. Statistics of ForgeryNIR. (a) The proportion of each perturbation applied in ForgeryNIR-rand in which the intensity of each perturbation ranges from zero to five. (b) The proportion of pertubations in ForgeryNIR-mix in which the intensity of each perturbation is fixed at one.

We set five intensity levels for each distortion. Fig. 4 shows example images selected from ForgeryNIR-rand in which the perturbation intensity ranges from 1 to 5. Considering the diversity and complexity of the real world, image data is often affected by more than one kind of distortion during transmission. Therefore, we construct a mixed subset, called **ForgeryNIR-mix**, in which the real and forgery face images are affected by more than one type of perturbation. The distribution of the number of perturbations used in the ForgeryNIR-mix dataset is shown in Fig. 5. To prevent the superposition of different perturbations from making the visual effect poor, we use the lowest perturbation intensity. The number of superimposed perturbations adopts a normal distribution

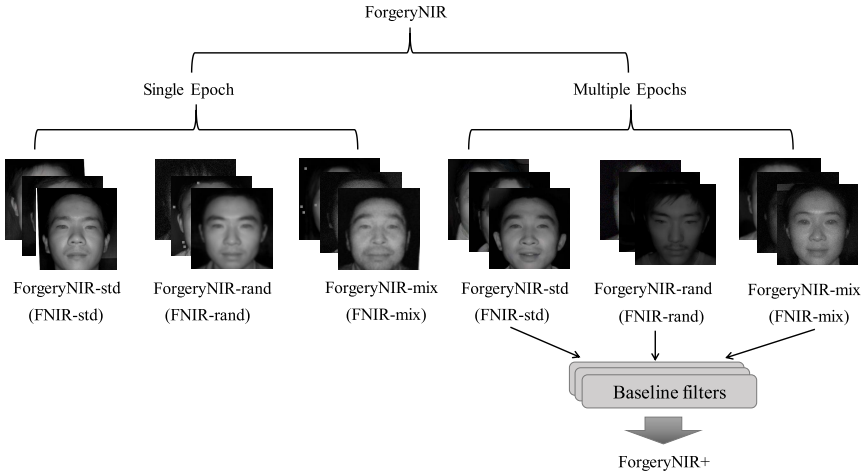


Fig. 6. Database composition and naming conventions. The suffix "std" in the naming represents directly generated faces without any perturbation, "rand" represents faces subject to only one random type of perturbation, and "mix" represents faces that are subject to more than one perturbation. The details are presented in section III.

method, with 2 types and 3 types being the majority, which is more representative of the real-world disturbance because extreme situations, *i.e.* extremely low distortion and extremely high distortion rarely occur in the real world. In addition, it is impossible for us to know the using forgery method in real-world scene. Inspired by [40], we also treat the images generated by the same type of GAN model saved under different epochs as different categories. Therefore, in order to enhance the generalization of the model, we use multiple epochs of four GANs (*i.e.*, cyclegan, progan, stylegan, stylegan2) while ensuring the quality of image generation. The images generated by different epoch models of the same GAN have no difference in visual observation. We use five quality evaluation metrics, including two early non-reference image quality assessment algorithms (niqe [22], brisque [23]) and three types of networks for neural image quality evaluation in NIMA [24] (mobilenet, nasnet, inception-resnet) to test the quality of fake images in our dataset ForgeryNIR-std. The quality evaluation experiment are presented in IV-B. For the face data generated by models which are saved under multiple epochs, we also provide perturbed versions. Database composition and naming conventions are shown in Fig. 6. The suffix **std** in the naming represents forgery faces that are without any perturbation, **rand** represents forgery faces that are subject to only one random type of perturbation, and **mix** represents forgery faces that is subject to more than one type of perturbation.

C. Hard Testing Set

Inspired by Deeperforensics-1.0 [12], the SOTA face forgery detectors are capable of almost ideal performance when training and testing sets are derived from the same dataset. This is not because the current forgery detection algorithm is powerful, but the existing datasets cannot contain all the forgery methods in the real world. It is just taken from a subset of all forgery methods. Moreover, attackers often conduct adversarial attacks based on open source face forgery detection algorithms. Therefore, in real-world scenarios, the models trained using these datasets may not perform a convincing evaluation due to the huge biases caused by a close distribution between the training and testing sets. In order to prevent attacks against samples, the testing set for real-world face forgery detection should contain as much data that the model is not easy to distinguish as possible, and the data distribution should also need to be different from the training set. A harder dataset is needed to better simulates complex real-world situations. It should satisfy three elements: (1) diversity (2) difficulty (3) quality.

Thus, in our paper, we further construct a harder testing set named ForgeryNIR+, which contains forged images that have bypassed existing face forgery detection methods. Firstly, we use four well-trained GANs (*i. e.* CycleGAN [2], ProGAN [3], StyleGAN [4], StyleGAN2 [5]) to generate faces in the near-infrared modality. Then, we added five common perturbations to the image to simulate the distortion in the real-world scene. Finally, we use three kinds of baselines (*i.e.*, [40]–[42]) to detect the testing set consisting of the clean data ForgeryNIR-std and the perturbated data ForgeryNIR-rand, ForgeryNIR-mix, and select the data of each kind of baseline misclassification as our hard testing set, named ForgeryNIR+. The hard testing set is more difficult to detect forged faces which contains images that are difficult to be distinguished from baselines. We further evaluate baselines performance on the challenging testing set ForgergNIR+ by devising some variants of the training set. Experiments show that accuracies of forgery detection methods on ForgeryNIR+ are only 10% to 30%, which shows much more difficulty. Therefore, our dataset not only creates a precedent for the subsequent research on forged faces in the near-infrared modality but also presents a more challenging ForgeryNIR+ subset which cannot be easily solved and be treated as a challenging benchmark for subsequent research.

D. Statistics of the Proposed Dataset

In order to better understand the ForgeryNIR and ForgeryNIR+ dataset, we conduct a statistical analysis of our dataset as follows.

TABLE III

IMAGE QUALITY ASSESSMENT (IQA) RESULTS OF FIVE METHODS. THE AVERAGE QUALITY SCORES $\mu \pm \sigma$ OF 1000 CLEAN IMAGES UNDER NIR MODALITAY GENERATED BY EACH GAN ARE PRESENTED RESPECTIVELY, WHERE μ AND σ REPRESENT MEAN AND STANDARD DEVIATION OF SCORE, RESPECTIVELY

Method	Real	CycleGAN	ProGAN	StyleGAN	StyleGAN2
Niqe [22]	4.70±0.795	7.00±0.612	5.26±0.725	5.56±0.695	6.06±0.674
Brisque [23]	37.62±7.782	52.51±2.426	39.89±7.914	45.03±7.616	48.72±6.106
Mobilenet [24]	4.64±0.034	4.22±0.041	4.75±0.035	4.94±0.012	4.83±0.024
Nasnet [24]	3.78±0.052	3.87±0.098	3.92±0.053	3.83±0.035	3.91±0.048
Inception-resnet [24]	3.96±0.028	4.45±0.043	3.98±0.024	4.06±0.021	4.03±0.022

1) ForgeryNIR: Most of the existing forgery detection datasets have no investigation on NIR modality and confine to address faces under VIS modality. However, the models trained using these datasets are not well adapted to the complex scenes of the real world. Thus, our presented dataset ForgeryNIR fills the gap of face forgery detection in the NIR modality. [40] treats the images generated by the same type of GAN model but saved under different epochs as different categories. Inspired by it, we construct two types of forged faces in our dataset. The image quality generated by these models is indistinguishable visually. In each setting, we use the four GANs (*i.e.*, CycleGAN, ProGAN, StyleGAN, and StyleGAN2) which is currently used in generating high-resolution images to generate 40,000 fake images, with a ratio of 1:1:1:1. We call this subset as ForgryNIR-std. Considering that images are often affected by signal distortion during transmission, so in order to better simulate real-world situations, we add 5 common perturbations to these clean faces. Types of perturbations are shown in Table II. Each disturbance is divided into 5 intensity levels. We added random perturbation types to 10,000 real images and 40,000 fake images. The distribution ratio of each disturbance type and disturbance intensity is shown in Fig. 5. Considering that the situtation is complex, and the image is often affected by more than one kind of perturbation in the real-world scenarios. Therefore, we further construct a subset called ForgreyNIR-mix in which each image is affected by 1-5 kinds of distortions. In terms of annotations, we not only provide true and false labels but also according to the GAN generation algorithm used to label the image by which method is generated, which provides convenience for subsequent research. The composition and naming conventions of the ForgeryNIR dataset are shown in Fig. 6.

2) ForgeryNIR+: The original intention of face forgery detection is to find a model with high accuracy and good generalization in rea-world scenes. Although the performance of some experiments conducted in previous works [40]–[42] is relatively high under several settings, these results are not convincing, because the training set and testing set used in these experiments have similar data distributions. Therefore, we propose a hard testing set to counter the challenge. On one hand, it avoids the problem of unreliable model accuracy due to the overly consistent distribution of the training set and the testing set. On the other hand, the images selected by ForgeryNIR+ are the forged faces that bypasses the baseline forgery detection filters, the risk of confrontational attacks is mitigated to a certain extent. First, we use the undisturbed clean data in ForgeryNIR, named ForgeryNIR-std, and the dataset ForgeryNIR-mix that is subject to more than one perturbation as the final testing set. The reason for this selection is that Forgery-std and ForgeryNIR-mix can imitate the forgery and disturbance in real-world to the greatest extent. Finally, we carefully selected 2,000 forged faces that bypass these threebaselines as our supplementary dataset, ForgeryNIR+.

E. Cross-Modality Knowledge Distillation

With the emergence of the new forgery generation technologies such as GAN, the spread of a large number of fake images and videos on the Internet has brought an extreme impact on society. A lot of research efforts have been dedicated to the field of face forgery detection, but most of these models are based on the VIS modality, and there is little involved in the forgery detection in the NIR modality. How to transfer the knowledge of the VIS model to the NIR modality is an urgent problem to be solved. Therefore, we propose a novel method based on knowledge distillation to avoid the problem of forgotten knowledge. The framework of our method is shown in Fig. 7. Both the teacher and student networks are selected as Xception [44]. First, we train a teacher model which can achieve an ideal accuracy in VIS modality using the dataset WildDeepfake [45] in which the real and fake videos are collected from the internet. Then we adopt a distillation strategy to train the student network. The accuracy of the pre-trained teacher model in the subset ForgeryNIR-std is only about 50%. This is reasonable because the model has never seen near-infrared data. Therefore, we update the teacher model at the same time during the distillation process, so that the teacher model not only retains the various forgery methods and perturbations in response to VIS modality but also has a certain generalization for NIR data. To the end, we propose a new constraint loss, which can be written as:

$$\mathcal{L}_{CMKD} = \alpha_1 \mathcal{L}_{CE_t} + \alpha_2 \mathcal{L}_{CE_s} + \alpha_3 \mathcal{L}_{KD} \tag{1}$$

where α_1, α_2, and α_3 are hyperparameters to control the three loss terms. According to experience, we set the α_1, α_2 to 2, and α_3 to 0.5. $\mathcal{L}_{CE_t}$ is the cross-entropy loss for the teacher network, $\mathcal{L}_{CE_s}$ is the cross-entropy loss for the student network. $\mathcal{L}_{KD}$ is the joint knowledge distillation loss of the teacher network and the student network which was first proposed in [46]. Since most of the current forgery detection models and the publicly available data sets used for face forgery

TABLE IV

BINARY CLASSIFICATION FACE FORGERY DETECTION ACCURACIES OF THE FORGERYNIR DATASET UNDER OUR PROPOSED CROSS-MODALITY KNOWLEDGE DISTILLATION (CMKD) METHOD AND SIX BASELINE FILTERS. WE EXPLORED THE TRAINING SET AND THE TESTING SET TO ADOPT DIFFERENT SETTINGS: STANDARD CLEAN DATA AND ADDING MORE THAN ONE TYPE OF PERTURBATION. ALL THE GENERATED IMAGES ARE UNDER MULTIPLE EPOCHS STRATEGY

Train	Test	GANFingerprint	CNNDetection	GANDCTAnalysis	Wavelet-Packet	Xception	MLP	CMKD
std	std	100%	100%	100%	100%	99.60%	63.08%	100%
	mix	62.52%	81.05%	65.34%	68.86%	64.12%	50.71%	72.72%
mix	mix	92.63%	96.57%	86.95%	87.34%	96.05%	58.33%	88.82%

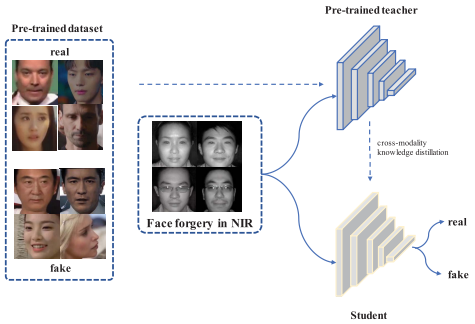


Fig. 7. The framework of our proposed face forgery detection method based on knowledge distillation. The teacher model is trained with the Xception [44]. The NIR domain data is provided to both teacher and student models. In addition to the student model, we update the teacher model at the same time during the distillation process, so that the teacher model not only retains the various forgery methods and perturbations in response to the VIS modality but also has a certain generalization for NIR data.

detection based on VIS modality are designed for authenticity detection and there is no traceability of the type of forgery, we perform a binary classification experiment based on our proposed method as well as the other baselines. The proposed cross-modality knowledge distillation (CMKD) method can achieve satisfactory performance on our novel topic of face forgery detection in NIR. Due to the time limitation, we merely apply Xception as our backbone model. The performance may be further improved when advanced networks are applied. The binary face forgery detection accuracies are shown in Table IV. When both the training set and the test set are selected as ForgeryNIR-std, in addition to [47], the other baselines and our proposed CMKD method have achieved ideal accuracies. So, this clue for high-frequency spectral decay discrepancies between the real images and the CNN-generated images in the VIS modality may not be applicable in the NIR modality. From our observation, when the training set uses ForgeryNIR-std and the testing set uses ForgeryNIR-rand, our method is higher than the other baselines. This means that the distillation strategy can improve the robustness of the student model when the testing set differs from the training set.

IV. EXPERIMENTS

In our experiments, we use four GANs mentioned above to generate 40,000 forgery images, and 10,000 real images are collected from CASIA NIR-VIS 2.0 as our proposed ForgeryNIR dataset. Each subset of the proposed dataset is divided into the training set, validation set, and testing set with a ratio of 7:1:2. Some other variants are mainly designed to explore the real performance of the face forgery detection baselines in the real world. We manipulate the complex scenes in the real world as much as possible by adding random disturbances to the testing set and select models with different epochs of the generation algorithm. For the sake of fairness, these experiments are all performed under the same dataset segmentation. Our experiment is similar to [40], not only detect the authenticity of a face image but also attribute the fake image to which GAN generated, which is conducive to fine-grained research in the future. The protocols of all the experiments on ForgeryNIR will also be released in the future.

A. Face Forgery Detection Baselines

The development of deep learning has brought new opportunities for problems of face forgery detection. We employ multiple current representative detection baselines, including GANFingerprint [40], CNNDetection [41], GANDCTAnalysis [42], Wavelet-Packet [48], Xception [49], and MLP [47] to evaluate our datasets ForgeryNIR and ForgeryNIR+ in detail. [42] found that the images generated by GAN in the frequency space have serious grid-like artifacts, and the upsampling operation caused by GAN shows that there are structural and fundamental problems in the way of generating images through GAN. [40] found that GAN models often leave stable fingerprint information in the process of generating images, and these pieces of fingerprint information are usually robust to different image frequencies and are not subject to image distortions and other influences. The original intention of [41] is to train a universal forgery detection model to distinguish real faces and images generated by CNN. And in order to verify the generalization ability of the model, a dataset composed of fake images generated by 11 different CNN-based image generator models was collected as a testing set, and the data was enhanced through careful preprocessing and postprocessing. Wavelet-Packet [48] is an improved method based on GANDCTAnalysis [42], which proposed a synthetic image detection model based on wavelet-packet representation of natural and GAN-generated images. [49] present a framework for evaluating detection methods under real-world conditions, consisting of cross-model, cross-data, and post-processing evaluation, and we evaluate their detection method to explore forgery detection on CNN-generated images using the proposed framework. [47] proposed the phenomenon that the high-frequency spectral components of CNN-generated images often do not decay as quickly as real images, and use this as a clue to perform CNN-generated faces detection. They concluded that this phenomenon was caused by the upsampling operation of GAN. We only selected one of their

TABLE V

FIVE CLASSIFICATION FACE FORGERY DETECTION ACCURACIES OF THE FORGERYNIR DATASET UNDER SIX BASELINE FILTERS. RESULTS OF (1) THE SINGLE EPOCH AND (2) MULTIPLE EPOCHS SETTINGS ARE PRESENTED. WE EXPLORED THE TRAINING SET AND THE TESTING SET TO ADOPT DIFFERENT SETTINGS: STANDARD CLEAN DATA SET, ADDING A RANDOM PERTURBATION AND ADDING MORE THAN ONE TYPE OF PERTURBATION. MOST OF THE PUBLIC AVAILABLE MODELS AND DATASETS USED FOR FACE FORGERY DETECTION BASED ON VIS MODALITY ARE DESIGNED FOR BINARY CLASSIFICATION. THERE ARE FEW OPEN SOURCES AND AVAILABLE PRE-TRAINED TEACHER MODELS UNDER VIS FOR OUR FIVE CLASSIFICATION TASKS. IN THIS EXPERIMENT, WE DID NOT COMPARE IT WITH THE CMKD METHOD

	Train	Test	GANFingerprint	CNNDetection	GANDCTAnalysis	Wavelet-Packet	Xception	MLP
(1)	std	std	99.93%	100%	100%	99.64%	99.97%	46.98%
		rand	56.93%	73.11%	39.23%	52.38%	56.14%	18.63%
		mix	59.56%	50.20%	30.43%	33.85%	46.31%	14.69%
	rand	rand	96.66%	98.25%	86.03%	87.87%	98.64%	26.70%
		mix	95.82%	92.04%	59.14%	69.51%	97.03%	29.18%
	mix	mix	98.14%	98.88%	70.94%	74.27%	98.90%	29.30%
(2)	std	std	99.92%	99.99%	99.99%	98.03%	99.19%	43.33%
		rand	52.80%	67.54%	27.80%	61.74%	59.88%	23.51%
		mix	52.33%	47.71%	20.58%	35.17%	51.51%	14.44%
	rand	rand	93.42%	97.20%	82.66%	87.37%	94.37%	27.17%
		mix	92.22%	91.01%	56.32%	69.17%	94.02%	28.85%
	mix	mix	94.76%	97.56%	64.82%	68.62%	96.15%	29.00%

model based on deep learning and applied it to our NIR dataset to explore whether the forged images in the NIR modality also conform to the pattern in the VIS modality.

B. Experimental Results on ForgeryNIR

Considering that the real-world scene is complex and uncontrollable, the data will often face the latent threat of distortion problems in the process of transmission, Therefore, our experiments mainly add the common perturbation methods to the standard dataset, and test the accuracy of the model under different settings. Furthermore, we consider the subsets in which each type of fake image is generated by carefully selected models saved under different epochs of the same GAN. In real-world scenarios, the forgery method selected by the forger and the number of training iterations of the model is uncontrolled, the fake videos on the Internet tend to show more diversity, so our multiple epochs experiments aim to explore the forgery images closer to the real-world scenarios. It can avoid to a certain extent the problem that the performance is remarkable due to the similar distribution of the training set and the testing set but is poor in real-world scenes.

1) IQA Experiments: We systematically conduct image quality assessment experiments on the fake images generated by each of GAN methods (CycleGAN, ProGAN, StyleGAN, StyleGAN2) in ForgeryNIR-std under multiple epochs setting. Five quality evaluation metrics we used include two early non-reference image quality assessment algorithms (*i.e.*, niqe [22], brisque [23]), where the smaller the value, the better the quality and 3 types of networks for neural image quality evaluation in [24] (*i.e.*, mobilenet, nasnet, inception-resnet), where the larger the value, the better the image quality. The IQA results are provided in Table III. In the nasnet and inception-resnet evaluation networks, the quality metircs of all the forged images are greater than the real images. The goal of NIMA [24] is to predict a higher correlation with human scores. Some generated images have higher scores than real images. This is consistent with the user study experiment. In addition, this is an objective method, so, it is reasonable for some generated images to have higher scores. However, in terms of quality assessment metrics based on machine learning, the results are different. The niqe and brisque quality evaluation metrics are based on the statistical law presented by the image spatial domain. These two metrics of all forgery faces are higher than those of real images, which means the quality of these forged images is inferior to real images. The niqe and brisque metrics of the images generated by ProGAN are 5.26 ± 0.725 and 39.89 ± 7.194 respectively, which are the closest to the real images. However, as the complexity of using GAN networks increases, the values of these two metrics are gradually reduced, that is, the quality of forged images becomes better.

2) Evaluation of the Effect of Perturbations: We evaluate the effect of perturbations on the performance of the forgery detection baselines. The results are shown in Table V. In all the settings, we used five distortion methods that are commonly used in the real world. We selected 10,000 real and 40,000 GAN-generated images with a total of 50,000 images as the standard subset. In terms of perturbation, we added random types of perturbation to ForgeryNIR-std, the intensity is set at random levels between level 0 and level 5. Inspired by [12], images are often affected by more than one kind of distortion. Therefore, a corresponding subset, named Forgery-mix, is constructed to better imitate the complexity and unpredictability in the real world. The division of these subset is the same as the standard dataset, with a ratio of 7:1:2. Macroscopically, no matter what variant settings we use for training and test sets, the detection accuracies of the last column [47] are significantly lower than other baselines. This is reasonable because the previous work [47], [49], [50] found that high-frequency spectral decay discrepancies

between the real and CNN-generated images. It is analyzed that this difference is caused by the upsampling operation that must occur in GAN. However, although there may be different up-sampling methods used, the discrepancies in the frequency spectrum between different types of forgery GANs are not obvious. Therefore, for the forgery GAN traceability experiment, the features extracted using this clue cannot distinguish the GAN and it is reasonable that the accuracies are low. Under the setting that the training set is selected as ForgeryNIR-std, when the testing set is also selected as clean data without any disturbance ForgeyNIR-std, the accuracy rate is almost 100%. It is reasonable because the training set and the testing set follow roughly the same distribution on the standard images without perturbations. When the testing set is selected as ForgeryNIR-rand and ForgeryNIR-mix, the accuracy of GANFingerprint [40] CNNDetection [41], Wavelet-Packet [48] and Xception [49] drop sharply and the accuracy of GANDCTAnalysis [42] is just 39.23% and 30.43% respectively on the two testing subsets. This phenomenon is reasonable because previous papers [36]–[39], have proposed that these disturbances can eliminate GAN fingerprints to a certain extent. The accuracy of GANDCTAnalysis [42] decreases further in the multiple epochs settings, which are just 27.80% and 20.58% on ForgeryNIR-rand and ForgeryNIR-mix respectively. When training and testing sets are selected from the same dataset, we find that SOTA face forgery detectors are capable of almost ideal performance in an ideal scenario. However, when detectors are evaluated on perturbed data which are not included during the training phase, the accuracy rates drop significantly. This shows that the perturbations destroy the frequency domain structure of the generated image to a certain extent, which makes the forgery detection model have problems in the attribution and traceability of the real and forgery images. Under the setting that the training set is selected as ForgeryNIR-rand, the results are obviously higher than the training set selected as ForgeryNIR-std. For GANFingerprint, when the testing sets are selected as ForgeryNIR-rand and ForgeryNIR-mix, the accuracy rate increased by 39.73% and 46.6%, and CNNDetection increased to 98.25% and 92.04%. Wavelet-Packet [48] as an improved method on GANDCT-Analysis, it can be seen that in most cases, whether the training set and testing set obey the same distribution, the accuracy is better than GANDCTAnalysis baseline. Although the accuracy rate of GANDCTAnalysis [42] is still the lowest, compared to the training set which has no perturbations, GANDCTAnalysis [42] increased by 46.8% and 28.71%, indicating that adding an appropriate amount of perturbation to the training data for data expansion can significantly improve the accuracy and make the face forgery detection model more robust to real-world scenarios. Therefore, the forgery detection model trained only with clean data cannot be well applied to the real-world scene, because the actual situation tends to be more diverse and complicated, and there will definitely be more than five types of perturbations we used in this paper. This requires that we should increase the diversity of the training set to better simulate the real-world distribution.

3) Evaluation of Multiple Epochs for NIR Face Forgery Detection: We conduct several evaluations of perturbations experiments on ForgeryNIR, but each type of fake images is generated based on the same GAN generation method, which also does not satisfy the distribution of data in real-world scenarios. GANFingerprint [40] treats the images generated by the same GAN model saved under different epochs as different categories. Inspired by the paper, we further consider the subsets in which each type of false image is generated by carefully selected models saved under different epochs of the same GAN. For the selection of multiple epochs, First, we select four GAN pre-training models, (*i.e.*, CycleGAN, ProGAN, StyleGAN, and StyleGAN2), each of which is trained based on the faces under VIS modality and can generate VIS face images with good visual effects. Then, we use each GAN model to perform finetuning on the real face under NIR modality in CASIA NIR-VIS 2.0. When the visual effect is already difficult to distinguish, we save the model every five epochs. After 180 epochs, a total of 36 models are saved, and these 36 models are used to generate our multiple epochs data. The experimental settings are the same as those in the setting of single epoch. The results are shown in Table V. For example, the GANFingerprint method has the lowest detection accuracy of 42.33% when the training and testing set are selected as ForgeryNIR-std and ForgeryNIR-mix respectively. Under the same conditions, the accuracy is 49.56% under the setting of single epoch. For GANDCTAnalysis baseline, the training and testing set are selected as ForgeryNIR-std and ForgeryNIR-mix respectively, the accuracy is only 20.58%, but it reaches 30.43% under the setting of single epoch. Comparing the results of single epoch, the accuracy of multiple epochs is reduced regardless of whether it is in the standard dataset or the dataset with perturbations.

TABLE VI

ABLATION EXPERIMENTS ABOUT FIVE TYPES OF PERTURBATIONS IN WHICH THE INTENSITY OF EACH PERTURBATION IS SET TO ONE. THE DETECTOR WE SELECTED IS CNNDETECTION. **GWN**: GAUSSIAN WHITE NOISE, **CC**: CHANGE IN COLOR CONTRAST, **JPEG**: JPEG COMPRESSION, **GB**: GAUSSIAN BLUR, **BD**: LOCAL BLOCK-BASED DISTORTION

GWN	CC	JPEG	GB	BD	Accuracy (%)
-	-	-	-	-	99.99
✓	-	-	-	-	20.43
-	✓	-	-	-	99.98
-	-	✓	-	-	99.92
-	-	-	✓	-	73.48
-	-	-	-	✓	40.43
✓	✓	-	-	-	20.42
✓	✓	✓	-	-	20.38
✓	✓	✓	✓	-	54.86
✓	✓	✓	✓	✓	34.73

4) Ablation Experiments: With the intention of exploring the impact of different types of perturbations on the performance of forgery detection algorithms, we conduct ablation experiments on five types of perturbations. And we sort out a table of ablation experiments based on the CNNDetection baseline, in which the training set is the subset ForgeryNIR-std without perturbations, and different perturbations were added

TABLE VII

FACE FORGERY DETECTION ACCURACIES OF THE FORGERYNIR+ DATASET IN THE BASELINES UNDER THE STANDARD SET WITHOUT PERTURBATIONS (FNIR-STD), STANDARD SET WITH RANDOM-LEVEL PERTURBATIONS (FNIR-RAND), STANDARD SET WITH MIXED PERTURBATIONS (FNIR-MIX), THE TRAINING SET IS UNDER MULTIPLE EPOCHS

	Train	Test	GANFingerprint	CNNDetection	GANDCTAnalysis
Multiple epochs	FNIR-std	FNIR-std	99.92%	99.99%	99.99%
		FNIR+	23.15%	36.95%	11.50%
	FNIR-std & FNIR-rand	FNIR-std & FNIR-rand	95.14%	98.56%	91.85%
		FNIR+	62.40%	56.70%	24.15%
	FNIR-std & FNIR-mix	FNIR-std & FNIR-mix	95.82%	99.21%	85.61%
		FNIR+	73.00%	86.05%	48.64%

to the testing set. We add five types of perturbation with the intensity of one to the testing set respectively. Based on the experimental results, we choose the disturbance with the greatest impact, that is, the lowest accuracy, as our initial perturbation. Then we add CC, JPEG, GB, BD perturbation according to the degree of impact from low to high. The experimental results are shown in Table VI. It can be seen that Gaussian white noise (GWN) has the greatest impact on the accuracy of the detection results, and the experimental result is only 20.43%, which is equivalent to no detection ability for the five-classification task. When changes in color contrast (CC) and JPEG compression (JPEG) are added in sequence, the accuracy is almost no change. This indicates that CC and JPEG have little effect on forgery detection in the near-infrared modality. The very strange point is that when we add Gaussian blur (GB), the accuracy has been greatly improved, from 20.38% to 54.86%. This may be due to Gaussian blur that weakens other disturbances to a certain extent.

C. Experimental Results on ForgeryNIR+

In the ForgeryNIR experiment, the test accuracy of the training set and the testing set is close to 100% when both the training set and the testing set are not added perturbations. In addition to distortions, forged images in the real world are often adversarial samples selected for forgery detection algorithms that are not included in our training set. From the evaluation of perturbations, we find the possibility of augmenting the training set to improve forgery detection model performance. Thus, we further evaluate baseline performance on the supplementary testing set by devising some variants of the training set. We perform experiments on ForgeryNIR+. In this setting, we use ForgeryNIR-std, ForgeryNIR-rand, and ForgeryNIR-mix under the setting of multiple epochs (with the same setting as the former experiments). Row 3 in Table VII shows the low accuracy when the models trained on ForgeryNIR-std and tested on ForgeryNIR+. Row 5 indicates that the accuracies of all the baseline models increase when trained on ForgeryNIR & ForgeryNIR-rand. In a more complex setting, when the models are trained on ForgeryNIR & ForgeryNIR-mix, row 7 shows the accuracy of all the detection baselines further increase. Results suggest that designing suitable training set variants has the potential ability to help increase face forgery detection accuracy, and applying various distortions to ensure the diversity of training set is necessary.

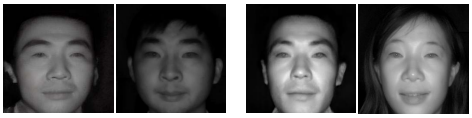


Fig. 8. Example of two images showing the lowest accuracy in the user study. The fake images are generated by the StyleGAN and ProGAN respectively, and the real images are randomly selected from CASIA NIR-VIS 2.0.

In row 3, 5, 7, the results are lower than the row 2, 4, 6. Therefore, in order to further explore the actual accuracy rate in the real-world scene, we construct the supplementary testing set, called ForgeryNIR+, containing forged images that have bypassed the face forgery detection methods.

D. User Study

Since it is difficult to measure the quality of the generated images in our proposed dataset ForgeryNIR quantitatively, we conducted a user study that can help evaluate the quality of the near-infrared face image generated using four GAN models. There are 71 subjects participated in the user study, and 57.75% of the subjects were students between 20 and 30 years old from Xidian University. Over half of the subjects did not have any prior experience in participation of user study and 56.34% of the subjects did not have any knowledge about the generative adversarial networks. The subjects were first asked to select their age, gender, and if they had any prior about the user study or the knowledge about GANs at first. Then, some real near-infrared face images taken from CASIA NIR-VIS 2.0 [43] are displayed in the questionnaire to help the users better judge the authenticity of the images. Finally, 30 groups of images were shown to the subjects, and each question showed one real image collected in CASIA NIR-VIS 2.0 with one image synthesized by CycleGAN, ProGAN, StyleGAN or StyleGAN2 with the four kinds of fake face images appeared with random chance. Participants were asked to select which one of the image is real. The average time taken for the questionnaire is around 4 minutes, which will not cause the effect of observer fatigue. We define each question to get 2 points for correct answers, and no points for incorrect answers. The maximum score is 60 points. The questionnaire of the anonymous user study is available online.

The score distribution results of the user study are shown in Fig. 9. It can be seen that over 70% of the participants scored below 40 of 60, and problems with accuracy less than 50%

TABLE VIII

FACE FORGERY DETECTION RESULTS ON THE IMAGES GENERATED BY DVG (PROVIDED BY THE AUTHORS). WE EXPLORED THE TRAINING SET AND THE TESTING SET TO ADOPT DIFFERENT SETTINGS: STANDARD CLEAN DATA SET, ADDING A RANDOM DISTORATION AND ADDING MORE THAN ONE TYPE OF DISTORATION

Train	Test	GANFingerprint	CNNDetection	GANDCTAnalysis
DVG/std	DVG/std	100%	100%	100%
	DVG/rand	98.55%	79.73%	51.64%
	DVG/mix	98.83%	64.35%	50.16%
DVG/rand	DVG/rand	100%	100%	97.15%
	DVG/mix	100%	99.9%	84.42%
DVG/mix	DVG/mix	100%	100%	95.32%

account for one-third of the total number of questionnaires, which demonstrates the good visual quality of the images in ForgeryNIR. The two most difficult images in this questionnaire are shown in Fig. 8, which are generated using ProGAN and StyleGAN2 respectively. It can be seen from the figure that the forged faces in our ForgeryNIR dataset are visually realistic compared with the real faces from CASIA NIR-VIS 2.0 dataset.

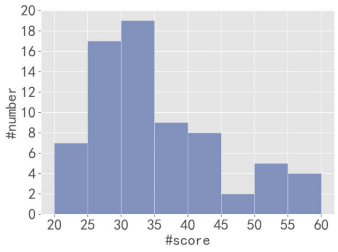


Fig. 9. Result of user study score distribution.

E. Discussions

Forged face generation in near-infrared modality can be applied to help boost the research of heterogeneous face recognition. For example, [21] is proposed to solve the problem of heterogeneous face recognition, which considers heterogeneous face recognition as a dual generation problem, and proposes a novel Dual Variational Generation (DVG) framework for NIR face generation. To better evaluate the performance of baselines on forged NIR faces, we additionally take an experiment based on 6,000 near-infrared face forged data generated by DVG provided by the authors. Furthermore, in our experiments we find that the real NIR images usually contain a certain amount of 0 pixels since their backgrounds are usually pure black. However, the numbers of 0 pixels in forged NIR images are usually very few. We summarize it as "0 pixel phenomenon" which will be discussed below. In addition, the perturbations we used are commonly used in face forged data, such as JPEG compression, Gaussian blur, Gaussian white noise, *etc*, and these perturbations are also widely used in current visible light-based face forgery datasets such as DeeperForensics-1.0 (2020) [12] and ForgeryNet (2021) [13]. In the pandemic time, we add a new perturbation like partial-face (due to face-mask) and make several experiments based on these faces wearing the mask.

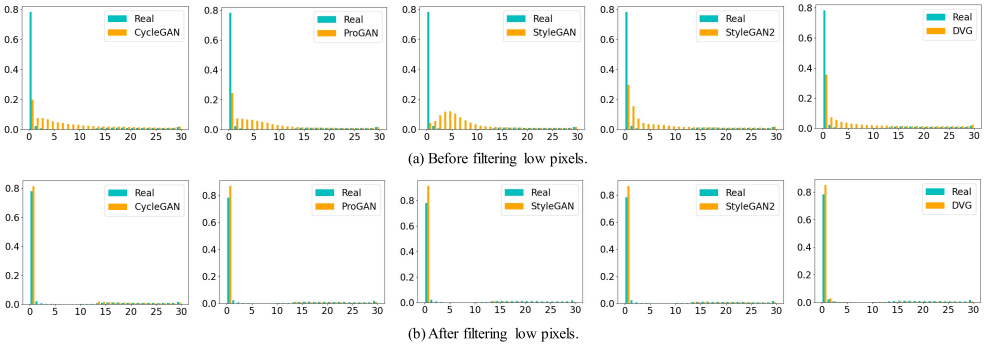
1) Experiment on DVG: We also conducted a face forgery detection experiment on 6,000 images with a resolution of 128×128 generated by DVG [21]. For the sake of fairness, The dataset is divided into the training set, validation set, and testing set with a ratio of 7:1:2, and use the same perturbation method as the previous experiments. The results of the experiment are presented in Table VIII. Face forgery detectors are capable of almost ideal performance when training and testing set are derived from the same dataset. For example, when both the training set and the testing set are selected as DVG/std, DVG/rand or DVG/mix, the detection results of the two baselines *i.e.*, GANFingerprint [39] and CNNDetection [41] reach the accuracy of 100%. For GANFingerprint baseline, when the training set only uses the subset without perturbations *i.e.*, DVG/std and the testing set uses the subset of DVG/rand with a random perturbation or the subset of DVG/mix with random number of disturbances, the accuracy is still close to 100%. This shows that GANFingerpint forgery detection algorithm is robust when the fake images are generated by DVG [21]. For CNNDetection, when the models are trained on DVG/std and tested on DVG/rand the accuracy drops by 22.27%. As the diversity of the testing set increases, the accuracy further decreases when the models are tested on DVG/mix. Perturbations destroy the frequency domain structure of the generated image to a certain extent, which makes the GANDCTAnalysis [42] method have problems in the attribution and traceability of the real and forgery images. Therefore, the accuracy of the GANDCTAnalysis baseline is the lowest compared with other detection algorithms.

2) "0 Pixel Phenomenon": In the process of analyzing the forged faces in NIR scenario, we found that the near-infrared face forged images generated by GANs have very different pixel distributions on the background compared to the real images. Real image backgrounds often show pure black, that is, all 0 pixels but GAN fake images tend to have a lot of low-pixel noise on the background, and relatively 0 pixel values are very few, which we called "0 pixel phenomenon". In addition, DVG [21] has generated some fake faces for heterogeneous face recognition research. We analyzed the fake faces generated by DVG and found this phenomenon too. The distribution of 0 pixels of real and fake images is shown in Fig. 10. The evaluation accuracy rate is shown in Table IX. We found that most of the detection accuracy rate after filtering

TABLE IX

FACE FORGERY DETECTION ACCURACIES OF THE FORGERYNIR DATASET BEFORE AND AFTER FILTERING OUT THE LOW PIXELS UNDER THREE BASELINES. WE EXPLORED THE TRAINING SET AND THE TESTING SET TO ADOPT DIFFERENT SETTINGS: STANDARD CLEAN DATA SET, ADDING A RANDOM PERTURBATION AND ADDING MORE THAN ONE TYPE OF PERTURBATION

	Train	Test	GANFingerprint		CNNDetection		GANDCTAnalysis	
			before	after	before	after	before	after
Multiple epochs	FNIR-std	FNIR-std	99.92%	99.69%	99.99%	100%	99.99%	99.99%
		FNIR-rand	52.80%	54.54%	67.54%	77.06%	27.80%	44.86%
		FNIR-mix	42.33%	34.08%	47.71%	50.41%	20.58%	24.91%
	FNIR-rand	FNIR-rand	93.42%	96.42%	97.20%	99.02%	82.66%	84.05 %
		FNIR-mix	92.22%	89.61%	91.01%	91.82%	56.32%	53.91%
	FNIR-mix	FNIR-mix	94.76%	96.90%	97.56%	98.94%	64.82%	67.48%



(a) Before filtering low pixels.

(b) After filtering low pixels.

Fig. 10. Comparison of different NIR face forgery images (generated by four GANs and DVG) with real NIR images on background pixel distribution. The first and the second rows are histograms of pixel distribution before and after filtering out low pixels. The blue columns represent the distribution of 0-30 pixel values of the real image background, and the orange columns represent the pixel distribution of the images generated by GANs or DVG, respectively.

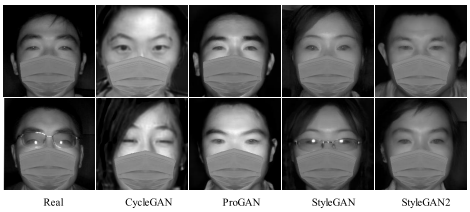


Fig. 11. Examples of perturbation of masks added to real and fake face images generated by four GANs in our ForgeryNIR datasets, where the real faces are collected from CASIA NIR-VIS 2.0.

out low pixels did not decrease. But this does not rule out that some other models can use this clue to detect forged NIR faces, which can be a potential research direction in the future.

3) Experiment on Faces With the Mask: In the pandemic time, we add a new perturbation like partial-face (due to face-mask)[8] to the subset ForgeryNIR-std. Examples are shown in Fig. 11. In order to explore the impact of this new type of perturbation on the accuracy, we conduct five-classification experiments based on the CNNDetection baseline separately. The experimental results are shown in Table X. We find that CNNDetection is capable of almost ideal performance when training and testing sets are derived from the same dataset. However, In the single epoch setting, when detectors are evaluated on the masked faces which are not included during the training phase, there is a significant drop in accuracy. Under the setting of multiple epochs, the accuracy rate can reach 87.45%, which is 54.68% higher than the accuracy in the single epoch setting. This proves that the multiple epochs strategy is efficient and effective to mask perturbation in the NIR modality.

[8]The code is available in https://github.com/aqeelanwar/MaskTheFace.

TABLE X

FACE FORGERY DETECTION ACCURACIES OF THE FORGERYNIR-STD SUBSET WITH AND WITHOUT THE MASK UNDER THE CNNDETECTION BASELINE FILTER. RESULTS OF THE SINGLE EPOCH AND MULTIPLE EPOCHS SETTINGS ARE PRESENTED. WE EXPLORED THE TRAINING SET AND THE TESTING SET TO ADOPT DIFFERENT SETTINGS: STANDARD CLEAN DATA AND ADDING THE MASK PERTURBATION

	Train	Test	Accuracy
Single epoch	FNIR-std	FNIR-std	100%
		FNIR-std (mask)	32.77%
	FNIR-std (mask)	FNIR-std (mask)	99.99%
Multiple epochs	FNIR-std	FNIR-std	99.99%
		FNIR-std (mask)	87.45%
	FNIR-std (mask)	FNIR-std (mask)	99.99%

4) Discussion of Face Forgery in NIR Compared With VIS Spectrum: NIR and VIS are different in image generation mechanism, so there are researches specifically on NIR face domain tasks. [19] pointed out that due to the difference of

spectral components between NIR and VIS modalities, visible light and near-infrared (VIS-NIR) face recognition is still a challenging task. Furthermore, the NIR-based face recognition systems are also threatened by face forgery attacks. In the frequency domain, we also found high-frequency spectral decay discrepancies. Unlike images in the VIS modality, the real images in the near-infrared modality and generated by GANs have no obvious difference in the decay of high-frequency components. There are two possible reasons: (1) the up-sampling operation in the GAN proposed by [47], [49] will cause the high-frequency components to be difficult to attenuate, which is not obvious in the near-infrared modality; (2) due to the different generation mechanisms of the near-infrared and the visible light images, the decay of the high-frequency components of the real images in the near-infrared modality may not be obvious. Therefore, this forgery detection method found in VIS is not applicable as a forgery detection clue in the NIR modality.

V. Conclusion

In this paper, we present an attempt at constructing a large-scale dataset named ForgeryNIR for face forgery detection in near-infrared (NIR) modality and porpose a novel method, named cross-modality knowledge distillation which achieve satisfactory performance on our novel topic of face forgery detection in NIR. Based on this data set, we then performed baseline evaluations on the representative forgery detection methods and analyzed the performance of these benchmarks on our ForgeryNIR dataset as well as the harder ForgeryNIR+ subset. This is a research concerning the problem of face forgery in the near-infrared modality. To better simulate the real-world scene, we provide a harder testing set containing forged images that have bypassed the face forgery detection method.

In this paper, we only constructed an image-based forgery detection dataset in the near-infrared modality. In real life, forged videos are more common. Therefore, this paper is also an inspiration for the future video-based near-infrared forgery detection research. We will further investigate the use of the "0 pixel phenomenon" for NIR face forgery detection. Because the real-world scenario tends to be more diverse and complicated, there will definitely be more than five types of perturbations we used in this paper. We will explore more perturbations encountered in real-world situation. We will also investigate the performance of adversarial examples in NIR face forgery detection scenario in the future.

References

[1] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

[6] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. CVPR Workshops*, vol. 1, Jun. 2019, pp. 1–8.

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

[8] 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.

[9] (2019). *Fakeapp*. [Online]. Available: https://www.fakeapp.com/

[10] 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.

[11] 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.

[12] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2889–2898.

[13] Y. He *et al.*, "ForgeryNet: A versatile benchmark for comprehensive forgery analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4360–4369.

[14] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5778–5788.

[15] J. Pu *et al.*, "Deepfake videos in the wild: Analysis and detection," in *Proc. Web Conf.*, Apr. 2021, pp. 981–992.

[16] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "KoDF: A large-scale Korean DeepFake detection dataset," 2021, *arXiv:2103.10094*.

[17] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face presentation attack with latex masks in multispectral videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 81–89.

[18] 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.

[19] H. Wang, H. Zhang, L. Yu, L. Wang, and X. Yang, "Facial feature embedded CycleGAN for VIS-NIR translation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1903–1907.

[20] J.-S. Park, D.-K. Hyun, J.-U. Hou, D.-G. Kim, and H.-K. Lee, "Detecting digital image forgery in near-infrared image of CCTV," *Multimedia Tools Appl.*, vol. 76, no. 14, pp. 15817–15838, Jul. 2017.

[21] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "Dual variational generation for low-shot heterogeneous face recognition," in *Proc. NeurIPS*, 2019, pp. 1–10.

[22] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[23] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[24] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[25] (2019). *Deepfakes*. [Online]. Available: https://github.com/deepfakes/faceswap/

[26] 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.

[27] (2019). *FaceSwap*. [Online]. Available: https://github.com/MarekKowalski/FaceSwap/

[28] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, 2019.

[29] (2019). *DeepFaceLab*. [Online]. Available: https://github.com/iperov/DeepFaceLab/

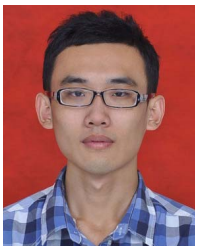
[30] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7184–7193.

[31] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

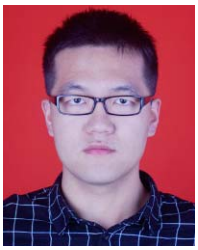
[32] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, Apr. 2015.

[33] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1818–1830, Aug. 2016.
[34] R. Wang *et al.*, "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," 2019, *arXiv:1909.06122*.
[35] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 15, 2020, doi: 10.1109/TPAMI.2020.3009287.
[36] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.
[37] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proenca, and J. Fierrez, "GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1038–1048, Aug. 2020.
[38] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-generated fake images over social networks," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 384–389.
[39] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?" in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Mar. 2019, pp. 506–511.
[40] 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.
[41] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot…for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8695–8704.
[42] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3247–3258.
[43] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 348–353.
[44] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[45] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2382–2390.
[46] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[47] K. Chandrasegaran, N.-T. Tran, and N.-M. Cheung, "A closer look at Fourier spectrum discrepancies for CNN-generated images detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7200–7209.
[48] M. Wolter *et al.*, "Wavelet-packet powered deepfake image detection," 2021, *arXiv:2106.09369*.
[49] N. Hulzebosch, S. Ibrahimi, and M. Worring, "Detecting CNN-generated facial images in real-world scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 642–643.
[50] T. Dzanic, K. Shah, and F. Witherden, "Fourier spectrum discrepancies in deep network generated images," 2019, *arXiv:1911.06465*.

Yukai Wang received the B.E. degree in computer science and technology from Hebei Normal University, Hebei, China, in 2020. He is currently pursuing the master's degree with Xidian University. His current research interests include computer vision, pattern recognition, and machine learning.



Chunlei Peng (Member, IEEE) received the B.Sc. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2012, and the Ph.D. degree in information and telecommunications engineering in 2017. From September 2016 to September 2017, he has been a Visiting Ph.D. Student with Duke University, Durham, NC, USA. He currently works with the School of Cyber Engineering, Xidian University. His current research interests include computer vision, pattern recognition, and machine learning.



Decheng Liu received the B.Sc. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2016, and the Ph.D. degree in information and telecommunications engineering in 2021. He currently works with the School of Cyber Engineering, Xidian University. His current research interests include computer vision and machine learning, especially for heterogeneous image analysis and its application.

Nannan Wang (Member, IEEE) received the B.Sc. degree in information and computation science from the Xi'an University of Posts and Telecommunications in 2009, and the Ph.D. degree in information and telecommunications engineering in 2015. From September 2011 to September 2013, he has been a Visiting Ph.D. Student with the University of Technology, Sydney, NSW, Australia. He currently works with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published more than 50 papers in refereed journals and proceedings, including *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His current research interests include computer vision, pattern recognition, and machine learning.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow at the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is also a Cheung Kong Professor of the Ministry of Education of China, a Professor of Pattern Recognition and Intelligent System with Xidian University, and a Professor of Computer Science and Technology with the Chongqing University of Posts and Telecommunications, Chongqing, China. He has published five books and around 200 technical articles in refereed journals and proceedings. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He is also a fellow of the Institution of Engineering and Technology. He served as the general chair/co-chair, a program committee chair/co-chair, or a PC member for around 30 major international conferences. He is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier).