TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 18/20 pp1839–1850
DOI: 10.26599/TST.2024.9010022
Volume 29, Number 6, December 2024

IIN-FFD: Intra-Inter Network for Face Forgery Detection

Qihua Zhou, Zhili Zhou*, Zhipeng Bao, Weina Niu, and Yuling Liu

Abstract: Since different kinds of face forgeries leave similar forgery traces in videos, learning the common features from different kinds of forged faces would achieve promising generalization ability of forgery detection. Therefore, to accurately detect known forgeries while ensuring high generalization ability of detecting unknown forgeries, we propose an intra-inter network (IIN) for face forgery detection (FFD) in videos with continual learning. The proposed IIN mainly consists of three modules, i.e., intra-module, inter-module, and forged trace masking module (FTMM). Specifically, the intra-module is trained for each kind of face forgeries by supervised learning to extract special features, while the inter-module is trained by self-supervised learning to extract the common features. As a result, the common and special features of the different forgeries are decoupled by the two feature learning modules, and then the decoupled common features can be utlized to achieve high generalization ability for FFD. Moreover, the FTMM is deployed for contrastive learning to further improve detection accuracy. The experimental results on FaceForensic++ dataset demonstrate that the proposed IIN outperforms the state-of-the-arts in FFD. Also, the generalization ability of the IIN verified on DFDC and Celeb-DF datasets demonstrates that the proposed IIN significantly improves the generalization ability for FFD.

Key words: deep learning; information security; image classfication; neural networks; face forgery; face forgery detection

1 Introduction

Recently, the face forgery technique has raised a security issue in that the human faces can be easily tampered and replaced to spread malicious information[1]. This face forgery technique is able to produce fake faces, which would lead to identity fraud issues in popular applications such as digital payment, video surveillance, and video call. Accordingly, there is an urgent need to explore face forgery detection (FFD) algorithms to detect whether the faces distributed on networks are fake ones to protect the identity information.

The existing approaches generally consider the FFD as a binary classification task that aims to distinguish real faces from fake ones. Some researchers produced datasets containing multiple forgery methods and adopt deep learning-based detectors to distinguish between the real and the fake ones[2–4]. Some researchers utilize generative models to obtain fake faces and detect them with deep learning-based detectors[5, 6]. CNN-based methods are also commonly applied to detect the suspect regions of the target image[7, 8]. However, these

• Qihua Zhou and Zhipeng Bao are with School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China. E-mail: zqh_1999_02_26@163.com; alexbao0206@outlook.com.
• Zhili Zhou is with Institute of Artificial Intelligence, Guangzhou University, Guangzhou 510006, China. E-mail: zhou_zhili@163.com.
• Weina Niu is with School of Computer Science, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China. E-mail: vinusniu@uestc.edu.cn.
• Yuling Liu is with School of Computer Science, Hunan University, Changsha 410082, Hunan, China. E-mail: yuling_liu@hnu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2023-11-16; revised: 2023-12-28; accepted: 2024-01-21

approaches generally focus on detecting the fake faces in some known datasets consisting of the fake faces produced by known forgery methods. Consequently, they are only able to detect fake faces in these known datasets, and thus show limited generalization ability of detecting the fake faces produced by unknown forgery methods. That makes them less appealing in practice. In practice, new kinds of face forgeries are emerged continuously, and thus the forgery detection approaches should achieve high generalization ability to resist unknown forgeries. Therefore, it is very necessary to improve the generalization ability to unknown forgeries.

In the literature, many approaches have been proposed to improve the generalization ability to unknown forgeries. For example, some recent approaches[9, 10] use the data augmentation strategy to improve the generalization ability for FFD. Although data augmentation can improve generalization ability to some known kinds of forgeries in the augmented data, it cannot handle more unknown kinds of forgeries. Meanwhile, some approaches[11, 12] learn intrinsic forgery features to improve the generalization ability. However, the difference between the data generated by a variety of forgeries is quite large, leading to poor generalization ability.

In this paper, we propose an intra-inter network (IIN) with a novel continual learning strategy to improve the generalization ability of FFD. Specifically, we decouple the common and special features of the different forgeries by the two feature learning modules. In particular, the intra-module is designed to learn special features by supervised learning and the inter-module is designed to learn the common features by self-supervised learning. The inter-module is an auxiliary module to optimize the training of the intra-module. The above feature decoupling can find the commonality between different features for forgery detection. Moreover, we design a forged trace masking module (FTMM) to mask the highly suspected parts of the forged faces and use it in contrastive learning to improve detection accuracy. The process of IIN is shown in Fig. 1. Our contributions can be concluded as follows:

- We propose an intra-inter network (IIN) to decouple common and special features in different forgeries. Then, we use two different modules to learn these two features. The inter-module is designed to learn the common features, while the intra-module is designed to learn the special features. Feature decoupling helps the IIN to focus on common features across the features of faces generated by different forgeries to improve the generalization ability for unknown forgeries.
- The self-supervised learning strategy, i.e., contrastive learning, is deployed to obtain superior common features to optimize the learning of special features. Therefore the generalization ability of IIN can be improved for unknown forgeries.
- A forged trace masking module (FTMM) is designed. It masks some regions of the face images that are more likely to be forged, and uses the images before and after masking as two views for contrastive learning. Maximizing the similarity between the input image and the masked image allows the IIN to focus on the forgery region, which is beneficial to the detection accuracy.

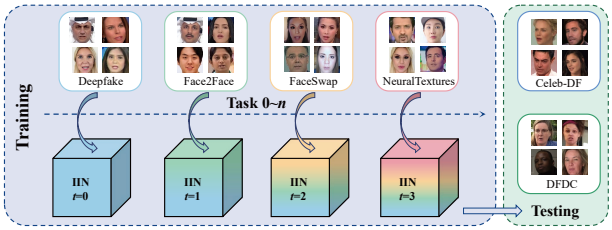

Fig. 1 Training of IIN is individually implemented on four training datasets, which are generated by four kinds of forgeries (i.e., Deepfake, Face2Face, FaceSwap, and NeuralTextures), respectively. The parameter t is deployed to control the order of training. The generalization ability of trained IIN is tested on unknown datasets (i.e., Celeb-DF and DFDC).

2 Related Work

2.1 Face forgery approaches

Early researches on face forgery perform face swapping and expression migration using simple operations based on neural networks. Korshunova et al.[13] proposed the first face swapping approach for replacing a target face with an input face, which replaces only the identity (i.e., information to identify a person) but keeps the expression unchanged. Recently, Faceshifter[14] has been proposed to generate high-fidelity forgery faces by integrating identity embedding of the input face and the attribute embedding of the target face.

Compared to face swapping, face reenactment produces more realistic faces and is therefore more

threatening. Face2Face[15] and NeuralTexture[16] were proposed to perform smooth expression manipulation. Face2Face developed a dense photometric consistency measure for facial expression tracking and achieved reenactment through effective deformation transfer. NeuralTexture has achieved higher face fidelity using neural networks. Fried et al.[17] tried to modify the talking-head video to manipulate the speech content smoothly. Suwajanakorn et al.[18] realized the transformation from audio to video, and successfully constructed the mouth shape and texture of the input audio to generate the forgery video.

Overall, these face forgery techniques have been able to produce high-fidelity forged faces to threaten social security. Therefore, the research on FFD is urgent.

2.2 Face forgery detection

Recently, many approaches have been proposed for FFD and made great progress[19, 20]. Early approaches mostly extract low-level features as classification clues to distinguish real faces from forgery ones. Li et al.[21] detected forgery artifacts by comparing texture differences around forged face boundaries. In some novel approaches[22, 23], high-frequency details are explored additionally for FFD. Gu et al.[24] adopted both the RGB and fine-grained frequency clues for FFD. Although these approaches can achieve superior accuracy in the known forgeries, the extracted low-level features limits their generalization ability to the unknown forgeries.

To improve the generalization ability, Li and Lyu[25] concluded that most existing face forgery approaches leave the common trace of blending an altered face into an existing background image, which is the key for generalization improvement. Liu et al.[26] adopted the phase spectrum to detect common artifacts of face forgeries for FFD. Miao et al.[27] captured fine-grained forgery details in the spatial and frequency domains to improve the generalization ability. Although these approaches attempt to capture the common features among different kinds of forgeries, they fail to decouple the common and special features, limiting their accuracy, and generalization.

2.3 Continual learning

The continual learning technique aims to learn certain parameters by employing a memory to store information of previous tasks, which can be broadly categorized into two groups, i.e., regularization-based approaches and experience replay approaches. In regularization-based approaches[28, 29], memory stores previously important parameters and avoids modifying crucial parameters of previous tasks when training a new task. In experience replay approaches[30, 31], partial samples of the previous task are stored in memory and applied to the process of training a new task.

Recently, some researchers try to design deep fake detection algorithms using continual learning. Khan and Dai[32] proposed a video transformer with face UV texture map for deepfake detection to improve detection accuracy. Kim et al.[33] designed a continual learning framework with knowledge distillation and apply it to deep fake detection. In summary, we argue that most of the existing approaches fail to decouple the common and special features across a sequence of tasks. Different from those approaches, we propose the IIN to decouple the common and special features with continual learning to resolve the generalization problem of FFD.

3 Approach

In this paper, we propose an IIN, which learns both the common features and special features from different tasks. That allows it to achieve higher accuracy and generalization in a continual learning framework. The proposed IIN mainly includes the two modules, i.e., intra-module and inter-module. (1) The inter-module is dedicated to learning common, task-agnostic features for guiding the intra-module; (2) the intra-module learns special, task-accessable features and consolidates the knowledge from the inter-module by using the labelled data in the current task.

The two modules of the IIN work in parallel. First, the inter-module learns the unlabeled samples in a fixed size memory K to optimize self-supervised learning (SSL) loss. At the same time, the inter-module learns the common features, which can also help the intra-module optimize the learning of the special features. Second, the objective of the intra-module is to optimise a supervised learning (SL) loss using labelled samples in the current task. Due to the general knowledge gained from the inter-module, the intra-module can adapt faster to the coming samples for higher accuracy. Moreover, the encoders of both modules have the same structure and the features learned by both sides can be easily fused for feature interaction.

3.1 Forged trace masking module

Existing forgeries methods generally adopt manual operations or neural networks to generate fake faces. The quality of different regions of generated fake face is different, and lower quality regions have higher possibility of being fake regions. To efficiently and accurately detect the forgery regions, it is reasonable for the neural network to focus more on the lower quality regions, i.e., forgery regions with higher suspicion. Therefore, in the proposed approach, we select the forgery regions with high suspicion for learning the detection network.

The main idea of the forged trace masking module is to mask the high suspected regions and make the masked image closer to the original for contrastive learning. The FTMM is shown in Fig. 2. We take a pre-trained extractor to calculate the gradients of the original image and select the highest gradient regions to mask them. The masking process can be described as follows:

$$I_{\mathrm{m}} = r \cdot I_{\mathrm{o}} \tag{1}$$

where r represents the forgery suspicion regions, I_{o} and I_{m} represent the original image and masked image, respectively. The masked and original images are given as two views for contrastive learning. The only difference between the two views is the masked regions, and making them closer allows the network to pay more attention to the masked regions to achieve higher classification accuracy.

3.2 Inter-module

3.2.1 Structure

The inter-module is responsible for capturing the common features between the different forgery tasks and using them to guide the learning of intra-module. As shown in Fig. 3, the inter-module receives two different views for contrastive learning. The first view V_a is the original image provided by memory K and the second view V_b is the masked image from FTMM. Since samples of past tasks are stored in memory allows the model to learn new knowledge while reviewing previous knowledge to acquire connections between tasks.

3.2.2 Loss function

The obtained two views are applied to optimize a contrastive loss L_{inter} by training a self-supervised encoder. To minimize computational resources while obtaining common features, we choose the SimSiam loss[34] as our self-supervised optimization objective

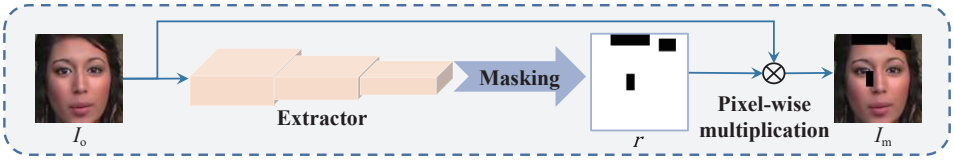

Fig. 2 Forged trace masking module (FTMM) uses an extractor to calculate the gradients of the original image for masking the highly suspected regions.

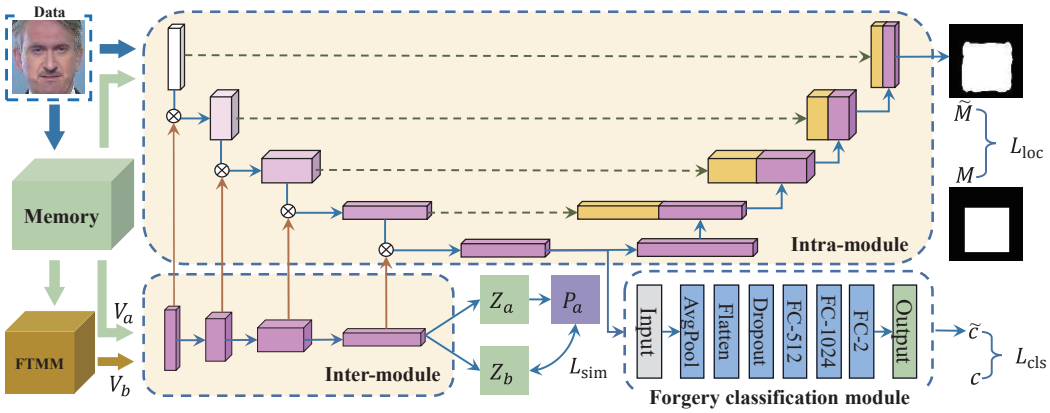

Fig. 3 Training of IIN, in which the memory mixes samples from the current and previous tasks, and feeds them to the IIN. FTMM calculates the highly suspected regions of samples from memory and masks them for contrastive learning. The inter-module and intra-module work simultaneously, and are linked by feature fusion. The final forgery features will be applied to both location detection and classification in FFD.

for the following advantages compared to the previous self-supervised loss: (1) for negative samples, no additional memory space is required (MOCO[35]), (2) no need for two duplicate networks (BYOL[36]), and (3) no requirement for handcrafted pretext loss (RotNet[37]). The two views V_a and V_b are fed to inter-module to obtain two features, i.e., Z_a and Z_b. Then, a prediction MLP head transforms the output of one view and compares it to the other view. The loss of inter-module is defined as

$$L_{\text{inter}} = \frac{1}{2}(D(P_a, \psi(Z_b)) + D(P_b, \psi(Z_a))) \tag{2}$$

where P is the output of the predictor, $\psi(\cdot)$ is the stop gradient operation, and $D(\cdot)$ is the negative cosine similarity,

$$D(P_a, Z_b) = -\frac{P_a}{\|P_a\|_2} \cdot \frac{Z_b}{\|Z_b\|_2} \tag{3}$$

where $\|\cdot\|_2$ is l_2-norm. SimSiam applies the same weights to construct two sets of features for comparison. Moreover, stopping the gradient propagation on one side of each set allows the network to implicitly alternate in updating the two sets of parameters.

3.3 Intra-module

3.3.1 Structure

To obtain both forgery location and classification, we choose UNet[38] as the architecture for the intra-module. As shown in Fig. 3, UNet extracts forgery feature in its encoder layers and produces the forgery location by concatenating the multi-scale features in its decoder layers. Previous works[39, 40] have demonstrated that both the low-level and high-level features are critical in FFD.

The intra-module is responsible for learning the special features for the current task under the guidance of the inter-module. To make the intra-module more easily adaptable to the feature guidance of the inter-module, we mix the coming data and memory data as training data for the intra-module. At the feature extraction stage, two structurally identical encoders fuse the feature maps of the corresponding layers. Specifically, let $\{c_i\}_{i=1}^{L}$ be the feature maps extracted from the inter-module's L layers on the image x. Correspondingly, intra-module utlizes L layers to extract features maps $\{s_i\}_{i=1}^{L}$. Feature fusion can be expressed as follows:

$$s_i = \xi_i(s'_{i-1}) \tag{4}$$

$$s'_i = s_i \otimes c_i, \quad i = (0, 1, ..., L) \tag{5}$$

where $\xi_i(\cdot)$ is the layer i of intra-module, s'_{i-1} and s'_i are the fused feature maps of layer $i-1$ and i, respectively. The final fused features s'_L can be obtained after L layers propagation. Extracted feature s'_L are applied for both forgery location detection and classification.

3.3.2 Loss function

For location detection, the decoder of UNet performs an up-sampling operation to produce a forgery mask $\tilde{M}$. We use Diceloss loss function has been applied to perform the detection of forgery location,

$$L_{\text{loc}} = 1 - \frac{2\left|M \cap \tilde{M}\right|}{|M| + \left|\tilde{M}\right|} \tag{6}$$

where M is the label mask and $\tilde{M}$ is the predicted forgery mask.

In the forgery classification, we apply a forgery classification module consisting of a set of simple fully connected layers to distinguish forgery faces. The binary cross entropy loss is adopted to optimize the parameters of the model,

$$L_{\text{cls}} = -(c \cdot \log(\tilde{c})) + (1 - c) \cdot \log(1 - \tilde{c}) \tag{7}$$

where c is the class label and $\tilde{c}$ is the predicted class label.

To recall knowledge of previous tasks, we create a mini memory m to store samples and labels $\{x, c\}$ for a past task h. We simultaneously train the data from the current task and a past task to achieve experience replay (ER). The experience replay loss can be expressed as follows:

$$L_{\text{exp}} = \frac{1}{|m|} \sum_{i=1}^{|m|} \text{BCE}(\tilde{c}_i, c_i) + \text{KL}(\tilde{c}_i \| \tilde{c}_h) \tag{8}$$

where BCE$(\cdot)$ is the binary cross entropy loss, KL$(\cdot)$ is the KL-divergence, and $\tilde{c}_h$ is the prediction of intra-module at the end of task h.

Three losses, L_{loc}, L_{cls}, and L_{exp} are adopted to train the intra-module, and we set three balance parameters to regulate the training process. During the training process, the balancing parameters will be automatically updated to achieve higher accuracy with the update of network parameters. In our experiments, we implement an effective automatic balancing strategy (ABS)[41] to optimize the balancing parameters. The final loss of the intra-module can be expressed as follows:

$$L_{\text{intra}} = \frac{1}{2\mu_1^2} L_{\text{loc}} + \frac{1}{2\mu_2^2} L_{\text{cls}} + \frac{1}{2\mu_3^2} L_{\text{exp}} + \log \mu_1 \mu_2 \mu_3 \tag{9}$$

where μ_1, μ_2, and μ_3 are optimizable parameters.

4 Experiment

Datasets: The FaceForensic++(FF++) is made up of 1000 real videos and 4000 forged videos constructed from four kinds of forgeries: DeepFake (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTexture (NT). Each kind of forgeries contains 1000 fake videos, which can be divided into three parts: 700 training videos, 200 validation videos, and 100 test videos. DFDC[42], Celeb-DF[43], DeeperForensics-1.0 (DF-1.0)[44], and FaceShifter (Shifter)[45] are adopted as unknown approaches to test generalization ability. There are 2500 real and 2500 fake videos in DFDC, 178 real and 340 fake videos in the Celeb-DF, 50 000 real and 10 000 fake videos in DF-1.0 and 1000 real and 1000 fake videos in Shifter. For each video we have taken images at 20 second intervals to make up the image datasets and all images are resized to 128×128.

Implementation details: The proposed IIN is trained on the NVIDIA GTX GeForce 3090 GPU platform, which is paired with 24 GB memory. All experiments were implemented using the Pytorch framework. We choose the UNet as the architecture for intra-module to detect forgery location. For the encoders of the IIN, we use the EfficientNet-b0 as the backbone network. The training process is performed end-to-end, updating the parameters of IIN for 10 epochs. We adopt Adam for our optimizer with the learning rate ranged from 10^{-2} to 10^{-6}. Moreover, the parameter μ_1, μ_2, and μ_3 are finally set to be 0.3015, 0.3561, and 0.3387 when the network converges stably. The predicted classification accuracy and location masks are reported. The link of the souce code is: https://github.com/QihuaZZ/AIMSGroup-IIN/.

4.1 Accuracy evaluation on FF++

With the rule of training one forgery for one task, the training order of the forgeries in FaceForensic++ is arranged as (1) DF, (2) F2F, (3) FS, and (4) NT. For different task identifier t, the forged location predictions for these forgeries are shown in Fig. 4. Moreover, the AUC evaluation for these four forgeries is shown in Fig. 5a.

As shown in Fig. 5a, it is clear to see that IIN performs the highest AUC on the forgery faces of the current task. That indicates that the model can easily achieve excellent performance in the intra-task condition. After training the first task, the model can hardly detect the forgery faces of the later tasks. However, when the training of the last task is completed, the performance of the model improved in the cross-task condition. The main reason for this is that the model replay experience from previous tasks to avoid catastrophic forgetting. Specifically, the inter-module receives the mixed data from memory K to extract the common features of different forgeries to guide the learning of the intra-module. Moreover, the experience replay loss L_{exp} is optimized by learning the data from mini memory m. Thus, IIN can learn the

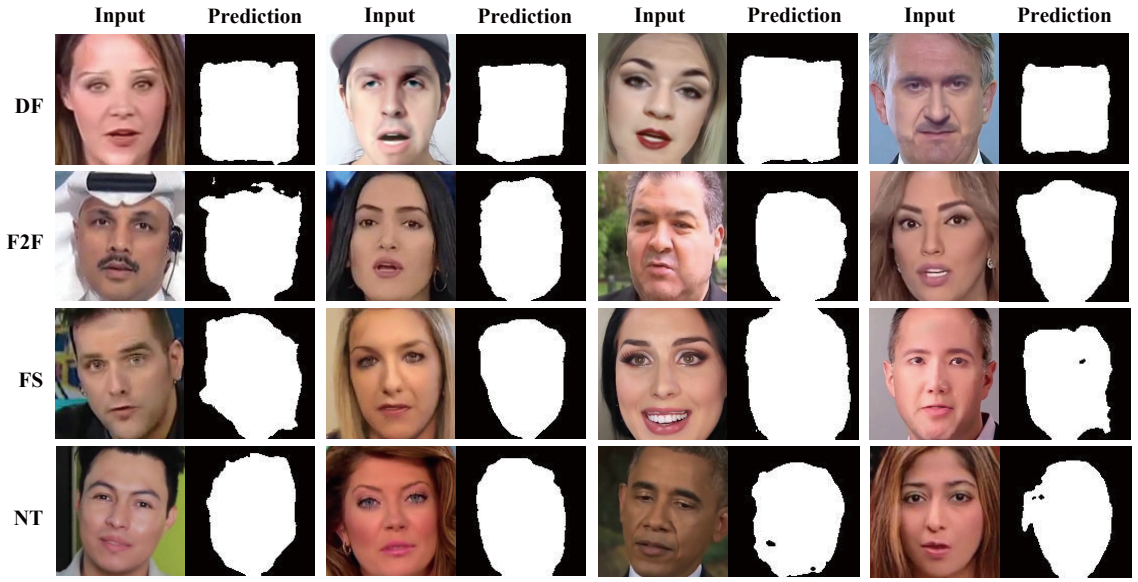

Fig. 4 Mask prediction by IIN.

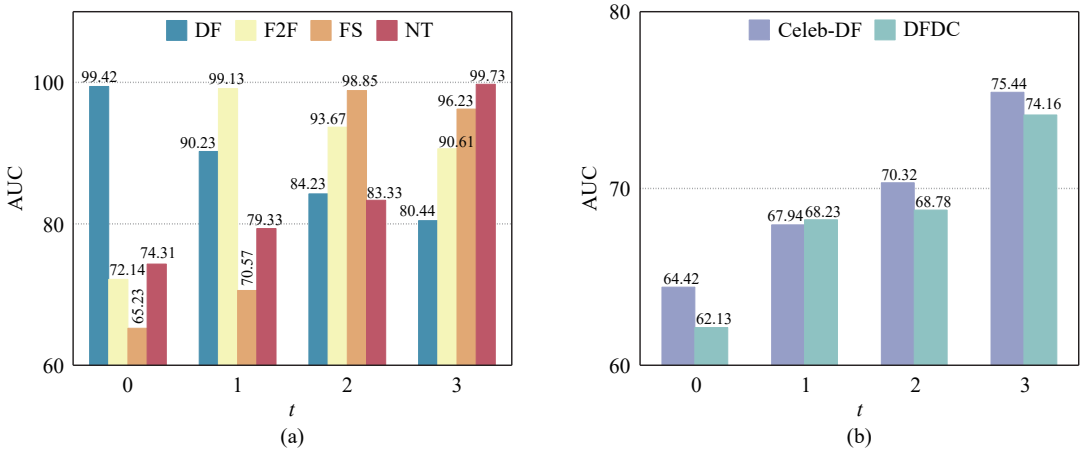

Fig. 5 (a) AUC evaluation on known datasets in FF++ at different task identifier t; (b) Generalization evaluation on unknown datasets at different task identifier t.

intrinsic commonality between different forgeries to avoid catastrophic forgetting.

4.2 Generalization evaluation on unknown forgeries

To evaluate the performance of the IIN on unknown forgery datasets, the Celeb-DF, DFDC, DF-1.0, and shifter are also employed to test the generalization ability for different task identifier t. The AUC values for the four datasets are shown in Fig. 5b.

From Fig. 5b, it is clear that the AUC values for the four datasets increase as the task identifier t changes. IIN requires the data from multiple forgeries to learn common and special features to obtain commonality between different kinds of forgeries to improve generalization ability. Thus, after completing the last task, the model gains strong inference, leading to high accuracy for unknown forgeries. Moreover, we compare the final generalization accuracy with the previous forgeries as shown in Table 1. Obviously, our approach achieves the highest detection accuracy in both the unknown forgery datasets.

4.3 Data distribution for datasets in FF++

To further demonstrate the commonality between the datasets in FF++. We adopt Adam[52] for our optimizer with the learning rate and apply t-distributed stochastic neighbor embedding (t-SNE)[53] to reduce the dimensionality of the samples of the four datasets in FF++ to observe the data distribution. From Fig. 6, the data distribution of the four datasets is highly coincided in a small region, which indicates that there could be commonality features among them. The learning of the commonality allows the model to focus more on the common features between the datasets, thus improving the generalization to unknown datasets

Table 1 Generalization comparisons with state of the arts.

(%)

Approach	Celeb-DF	DFDC	DF-1.0	Shifter
MesoInception[46]	50.24	49.87	50.58	51.36
Face artifacts[18]	57.32	60.27	58.74	60.48
Head pose[47]	54.60	56.18	58.29	56.30
Xception+Reg[48]	71.20	70.48	69.79	71.21
LRNet[49]	57.40	59.26	59.17	58.80
Xception[46]	65.50	67.40	68.39	66.76
MTD-Net[40]	70.12	72.38	71.42	70.70
Schwarcz chellappa[50]	67.44	67.30	68.13	68.42
Yu et al.[51]	74.20	70.09	67.14	72.18
Video transformer[32]	70.18	68.37	71.39	72.69
CoReD[33]	74.68	72.46	70.62	69.89
IIN	**75.44**	**74.16**	**73.39**	**74.69**

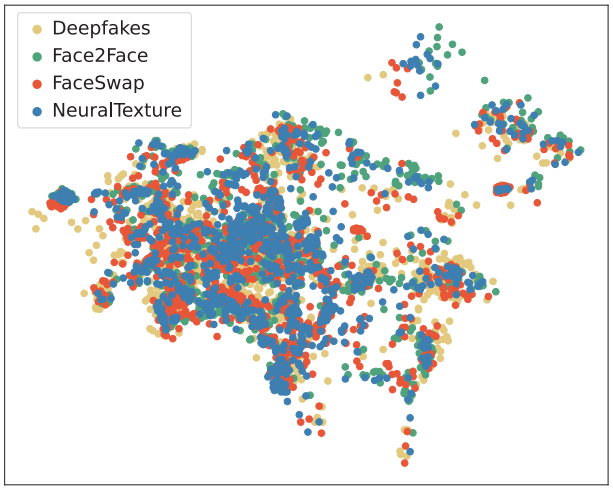

Fig. 6 Data distribution for datasets in FF++.

4.4 Robustness to unseen perturbations

Since images on social media are susceptible to processing, designing robust forgery detectors to resist common perturbations is necessary. After we train the IIN on the FF++ dataset, this trained detector is applied to detect FF++ samples with multiple perturbations to evaluate robustness. These perturbations include: changes in saturation, changes in contrast, adding blockwise distortions, adding white Gaussian noise, blurring, pixelating, and applying video compression. The robustness comparisons are shown in the Table 2.

In Table 2, it is clear that IIN outperforms other approaches in resisting multiple common perturbations. Obviously, the accuracy of methods except IIN decreases significantly when detecting images with perturbations affecting high-frequency content including blur, pixelation, and compression. Gaussian noise has the greatest impact on the prediction accuracy of the detector due to the fact that it severely disrupts key recognition features of the face. Moreover, both video transformer and CoReD are vulnerable to the block-wise distortions because they do not preprocess the input image making the trained model unable to cope with the perturbation. For IIN, the input image is subjected to FTMM to mask the high suspect region to achieve image enhancement, and then the masked image and the original image are fed into the inter-module for contrastive learning to obtain more comprehensive features. Thus, IIN can easily resist the block-wise distortions while maintaining high generalization.

4.5 Generalization evaluation of different training orders of datasets in FF++

In continual learning, multi-tasks are learned in a time order. During the training process of IIN, each task is responsible for training one dataset in FF++, and the training order of different datasets may affect the generalization ability.

We train the IIN in four different dataset orders and evaluate the generalization ability on Celeb-DF and DFDC with the trained models. From Table 3, it is clear that the model trained with four different dataset orders all achieve high generalization ability on both Celeb-DF and DFDC. That is because the memory stores the data from past tasks, and the inter-module can extract the commonality between these mixed data to improve the generalization ability. Thus, IIN can keep high generalization ability regardless of the training order of datasets.

We also evaluate the generalization ability of the IIN trained with the mixture of all face forgeries in FF++ without continual learning. Although its accuracy is similar compared to IIN trained with continual learning, it requires a mixed dataset of all face forgeries for each training process. However, the model trained by continual learning only needs to train a new

Table 2 Robustness to unseen perturbations.

(%)

Approach	Clean	Saturation	Contrast	Block	Noise	Blur	Pixel	Compress
Xception[46]	99.81	99.32	98.63	99.24	53.82	60.27	74.25	62.13
LRNet[49]	99.72	96.51	88.67	99.27	58.66	63.71	65.73	73.34
MTD-Net[40]	99.85	97.87	95.23	96.18	63.24	58.53	74.16	81.47
Yu et al.[51]	99.84	98.63	97.74	94.46	58.74	72.14	**89.67**	84.92
Video transformer[32]	99.66	87.64	97.89	78.12	65.43	68.98	69.86	69.89
CoReD[33]	99.82	89.29	98.19	83.25	68.38	72.87	76.57	75.44
IIN	**99.88**	**99.82**	**99.37**	**99.68**	**72.16**	**88.94**	83.89	**90.67**

Table 3 Generalization evaluation of four different training orders and the mixture of all face forgeries in FF++.

(%)

Test dataset	Training order																
	Order 1				Order 2				Order 3				Order 4				No order
	DF	F2F	FS	NT	F2F	FS	NT	DF	FS	NT	DF	F2F	NT	DF	F2F	DF	Mixture of all
Celeb-DF	75.44				73.72				74.63				75.16				73.46
DFDC	74.16				73.23				73.79				73.28				72.18
DF-1.0	74.47				73.68				73.12				72.15				73.74
Shifter	75.86				75.49				73.67				74.18				70.89

task on the new forgery dataset thus reducing the training cost. Face forgery algorithms progress continuously, and face forgery detection algorithms need to progress continuously accordingly, so a continuous learning framework is very applicable.

4.6 Ablation study

To gain a deeper understanding of IIN, we discuss three of its main factors that can affect accuracy, i.e., the inter-module for commonality learning, the FTMM for forged regions masking, and the ABS for parameter balancing.

4.6.1 Effect of inter-module and FTMM

The inter-module is designed for obtaining the common features by self-supervised learning. In the inter-module, unlabeled data is utilized for training to obtain comprehensive common features to guide the learning of the special features to improve the generalization ability. The FTMM is designed to mask highly suspected regions of the original image. Contrastive learning among the masked image and the original image is implemented so that IIN can pay more attention on the forged regions to improve the detection accuracy. Here we conduct the ablation experiment to prove their effectiveness. As listed in Table 4, the experimental results on known forgeries and unknown forgeries that both the Inter-module and FTMM have positive effect on our scheme.

4.6.2 Effect of ABS

The ABS is utilized to automatically select the balancing parameters $\mu 1$, $\mu 2$, and $\mu 3$. We train IIN with ABS for 10 epochs. From Fig. 7, our model achieves better generalization ability by implementing ABS.

4.7 Limitation

In this paper we propose an IIN by feature decoupling to improve the generalization ability. However, our approach still has the following limitation. Although we innovate on feature decoupling and achieve high generalization ability, the selection of training data has a significant impact on generalization ability. There is a large gap between the forgeries in different forgery datasets. If the training data comes from multiple datasets, there may not be similar modification traces between the different forgeries, so it is difficult for the network to extract their common features resulting in low generalization ability.

Nevertheless, we believe the proposed IIN can be a useful strategy to improve the generalization ability of FFD. Meanwhile, we hope that the shortcomings exposed by this work will contribute to the advancement of the field and give rise to innovative thinking.

5 Conclusion

In this paper, a novel FFD approach based on inter-intra network by continual learning is proposed to improve the generalization ability. First, the intra-module works as a backbone module to learn the special features of the forgery samples within the task to obtain both detection accuracy and forgery location. Meanwhile, considering that different forgeries may leave similar forgery traces, the inter-module is also designed to capture those common features across the tasks to improve the generalization ability of the IIN to unknown forgeries. These two modules work simultaneously and the inter-module guides the intra-module to learn superior forgery features by feature fusion. Moreover, the FTMM is designed to mask high suspected regions of the forgery faces sampled from the memory, the masked image, and the original image are fed together into the inter-module for contrastive learning, allowing the IIN to focus more on the forgery regions for higher detection accuracy. Extensive experiments show that our approach achieves high generalization ability to unknown forgeries as well as high detection accuracy to known forgeries.

Table 4 Ablation study of inter-module and FTMM.

(%)

Model setting			Dataset							
			Known forgery				Unknown forgery			
Intra-module	Inter-module	FTMM	DF	F2F	FS	NT	Celeb-DF	DFDC	DF-1.0	Shifter
√			96.74	96.03	97.24	96.19	72.36	71.19	69.43	70.84
√	√		97.31	96.87	96.39	96.58	75.44	74.16	73.39	74.69
√		√	99.16	99.48	98.84	99.73	71.61	72.25	69.44	70.16
√	√	√	99.23	99.19	99.13	99.28	75.37	74.68	74.12	73.86

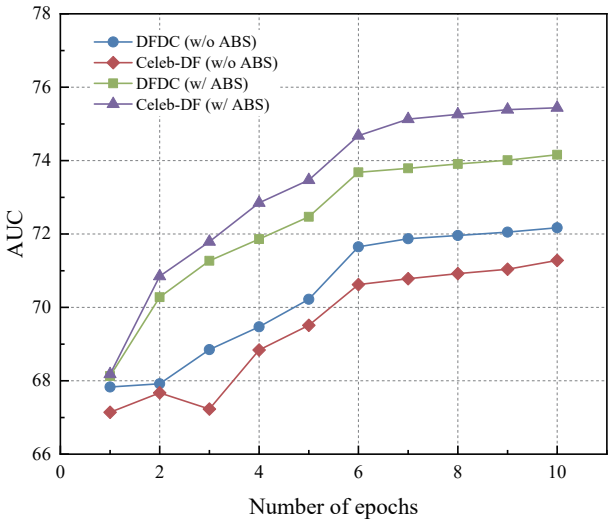

Fig. 7 Ablation study of the ABS.

References

[1] K. A. Pantserev, The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability, in *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, H. Jahankhani, S. Kendzierskyj, N. Chelvachandran, and J. Ibarra, eds. Cham, Switzerland: Springer, 2020, pp. 37–55.

[2] X. Ding, Z. Raziei, E. C. Larson, E. V. Olinick, P. Krueger, and M. Hahsler, Swapped face detection using deep learning and subjective assessment, *EURASIP J. Inf. Secur.*, vol. 2020, no. 1, p. 6, 2020.

[3] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niebner, Faceforensics: A large-scale video dataset for forgery detection in human faces, arXiv preprint arXiv: 1803.09179, 2018.

[4] 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. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 1–11.

[5] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, Detection of GAN-generated fake images over social networks, in *Proc. IEEE Conf. Multimedia Information Processing and Retrieval* (*MIPR*), Miami, FL, USA, 2018, pp. 384–389.

[6] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, Detecting both machine and human created fake face images in the wild, in *Proc. 2nd Int. Workshop on Multimedia Privacy and Security*, Toronto, Canada, 2018, pp. 81–87.

[7] J. Wang, S. C. Satapathy, S. Wang, and Y. Zhang, LCCNN: A lightweight customized CNN-based distance education app for COVID-19 recognition, *Mob. Netw. Appl.*, pp. 1–16, 2023.

[8] Y. D. Zhang, V. Govindaraj, and Z. Zhu, FECNet: A neural network and a mobile app for COVID-19 recognition, *Mob. Netw. Appl.*, pp. 1–14, 2023.

[9] X. Wei, S. Liang, N. Chen, and X. Cao, Transferable adversarial attacks for image and video object detection, arXiv preprint arXiv: 1811.12641, 2018.

[10] Z. Guo, G. Yang, J. Chen, and X. Sun, Fake face detection via adaptive manipulation traces extraction network, *Comput. Vis. Image Underst.*, vol. 204, p. 103170, 2021.

[11] X. Xuan, B. Peng, W. Wang, and J. Dong, On the generalization of GAN image forensics, in *Proc. Biometric Recognition*: *14th Chinese Conf., CCBR 2019*, Zhuzhou, China, 2019, pp. 134–141.

[12] M. Du, S. Pentyala, Y. Li, and X. Hu, Towards generalizable forgery detection with locality-aware autoencoder, arXiv preprint arXiv: 1909.05999, 2019.

[13] I. Korshunova, W. Shi, J. Dambre, and L. Theis, Fast face-swap using convolutional neural networks, in *Proc. IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 3677–3685.

[14] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, Advancing high fidelity identity swapping for forgery detection, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 5074–5083.

[15] 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. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 2387–2395.

[16] J. Thies, M. Zollhöfer, and M. Nießner, Deferred neural rendering: Image synthesis using neural textures, *ACM Trans. Graph.*, vol. 38, no. 4, p. 66.

[17] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, Text-based editing of talking-head video, *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–14, 2019.

[18] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, Synthesizing Obama: Learning lip sync from audio, *ACM Trans. Graph.*, vol. 36, no. 4, p. 95, 2017.

[19] X. Yang, Y. Li, and S. Lyu, Exposing deep fakes using inconsistent head poses, in *Proc. ICASSP 2019 - 2019 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, United Kingdom, 2019, pp. 8261–8265.

[20] Y. Li, M. C. Chang and S. Lyu, In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking, in *Proc. 2018 IEEE Int. Workshop on Information Forensics and Security* (*WIFS*), Hong Kong, China, 2018, pp. 11–13.

[21] Y. Li and S. Lyu, Exposing deep fake videos by detecting face warping artifacts, arXiv preprint arXiv: 1811.00656, 2018.

[22] Y. Luo, Y. Zhang, J. Yan, and W. Liu, Generalizing face forgery detection with high-frequency features, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 16317–16326.

[23] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues, in *Lecture Notes in Computer Science*, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, eds. 2020, vol. 12357, pp. 86–103.

[24] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi, Exploiting fine-grained face forgery clues via progressive enhancement learning, *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, pp. 735–743, 2022.

[25] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, Face X-ray for more general face forgery detection, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 5001–5010.

[26] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 772–781.

[27] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, F2Trans: High-frequency fine-grained transformer for face forgery detection, *IEEE Trans. Inform. Forensic Secur.*, vol. 18, pp. 1039–1051, 2023.

[28] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 13, pp. 3521–3526, 2017.

[29] F. Zenke, B. Poole, and S. Ganguli, Continual learning through synaptic intelligence, in *Proc. 34th Int. Conf. Machine Learning*, Sydney, Australia, 2017, pp. 3987–3995.

[30] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, Learning to learn without forgetting by maximizing transfer and minimizing interference, arXiv preprint arXiv: 1810.11910, 2018.

[31] Q. Pham, C. Liu, D. Sahoo, and H. Steven, Contextual transformation networks for online continual learning, in *Int. Conf. on Learning Representations*, Virtual Event, 2021.

[32] S. A. Khan and H. Dai, Video transformer for deepfake detection with incremental learning, in *Proc. 29th ACM Int. Conf. Multimedia*, Virtual Event, China, 2021, pp. 1821–1828.

[33] M. Kim, S. Tariq, and S. S. Woo, CoReD: Generalizing Fake Media Detection with Continual Representation using Distillation, in *Proc. 29th ACM Int. Conf. Multimedia*, Virtual Event, China, 2021, pp. 337–346.

[34] X. Chen and K. He, Exploring simple Siamese representation learning, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 15750–15758.

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, Momentum contrast for unsupervised visual representation learning, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 9729–9738.

[36] J. B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al., Bootstrap your own latent a new approach to self-supervised learning, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 21271–21284.

[37] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, Why does unsupervised pre-training help deep learning, in *Proc. thirteenth Int. Conf. on Artificial Intelligence and Statistics*, Sardinia, Italy, 2010, pp. 201–208.

[38] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Int. Conf. on Medical Image Computing and Computer-Assisted intervention*, Cham, Switzerland, 2015, pp. 234–241.

[39] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, DeepFake detection based on discrepancies between faces and their context, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6111–6121, 2022.

[40] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, MTD-net: Learning to detect deepfakes images by multi-scale texture difference, *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4234–4245, 2021.

[41] R. Cipolla, Y. Gal, and A. Kendall, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7482–7491.

[42] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, arXiv preprint arXiv: 2006.07397, 2020.

[43] 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. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 3207–3216.

[44] 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. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 2889–2898.

[45] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, Faceshifter: Towards high fidelity and occlusion aware face swapping, arXiv preprint arXiv: 1912.13457, 2019.

[46] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, MesoNet: A compact facial video forgery detection network, in *Proc. IEEE Int. Workshop on Information Forensics and Security* (*WIFS*), Hong Kong, China, 2018, pp. 1–7.

[47] R. Wang, J. F. Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces, arXiv preprint arXiv: 1909.06122, 2019.

[48] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, On the detection of digital face manipulation, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 5781–5790.

[49] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, Improving the efficiency and robustness of deepfakes detection through precise geometric features, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 3609–3618.

[50] S. Schwarcz and R. Chellappa, Finding facial forgery artifacts with parts-based detectors, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*

Workshops (*CVPRW*), Nashville, TN, USA, 2021, pp. 933– 942.

[51] P. Yu, J. Fei, Z. Xia, Z. Zhou, and J. Weng, Improving generalization by commonality learning in face forgery detection, *IEEE Trans. Inform. Forensic Secur.*, vol. 17, pp. 547–558, 2022.

[52] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv: 1412.6980, 2014.

[53] L. Van der Maaten and G. Hinton, Visualizing data using t-sne, *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.

Zhipeng Bao is a graduate student. He is currently pursuing the MS degree with Nanjing University of Information Science and Technology, China, in 2021. His research interest includes steganography.

Weina Niu received the bachelor degree in software engineering from Shenyang Normal University in 2011, and the PhD degree in computer software and theory from University of Electronic Science and Technology of China in 2018. She is currently an associate professor at School of Computer Science and Engineering, University of Electronic Science and Technology of China. Her research interests include malware analysis, network attack detection, and data security.

Yuling Liu received the PhD degree in computer application from Hunan University in 2008. She is currently an associated professor with College of Computer Science and Electronic Engineering, Hunan University. Also, she was a visiting scholar at UMASS Lowell. Her current research interests include multimedia security, artificial intelligence security, and information hiding.

Qihua Zhou is a graduate student. He is currently pursuing the MS degree in Nanjing University of Information Science and Technology, China, in 2021. His research interest includes steganography and deep fake detection.

Zhili Zhou received the MS and PhD degrees in computer application from Hunan University, in 2010 and 2014, respectively. He is currently a professor with Institute of Artificial Intelligence and Blockchain, Guangzhou University. Also, he was a postdoctoral fellow with Department of Electrical and Computer Engineering, University of Windsor, Canada. His current research interests include multimedia security, artificial intelligence security, information hiding, digital forensics, blockchain, and secret sharing. He has authored or co-authored more than 100 refereed papers. He is serving as an associate editor of *Journal of Real-Time Image Processing, Security and Communication Networks*, and *International Journal on Semantic Web and Information Systems*. He received ACM Rising Star Award and got Guangdong Natural Science Funds for Distinguished Young Scholar.