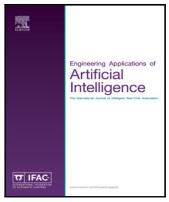
Engineering Applications of Artificial Intelligence 133 (2024) 108341

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Spatial-frequency feature fusion based deepfake detection through knowledge distillation

Bo Wang [a], Xiaohan Wu [a], Fei Wang [a], Yushu Zhang [b], Fei Wei [c], Zengren Song [d,*]

[a] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[b] School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[c] School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417, Singapore
[d] National Computer Network Emergency Response Technical Team/Coordination Center of China, Dalian 116021, China

ARTICLE INFO

Keywords:
Deepfake detection
Knowledge distillation
Frequency domain
Feature fusion

ABSTRACT

While the misuse of Deepfake technology is drawing growing concern in the literature of information security, related forgery detection has become a significant challenge in practical applications. Most state-of-the-art detection methods achieve satisfactory results on raw images, but their performance drops significantly on processed images (e.g. compression). In this work, we propose a novel Deepfake detection method that integrates spatial and frequency domain information within a knowledge distillation framework for efficient forgery detection. Our method consists of two steps: (1) spatial-frequency fusion, and (2) multi-knowledge distillation. We first extract frequency-domain and spatial-domain features, then fuse them and utilize them in attention-based guidance to improve the classification results. Note that the spatial-frequency fusion serves as the basis for both the teacher and student models with spatial-frequency features and logits transferred as knowledge. We conducted comprehensive experiments on several benchmark datasets which successfully demonstrate the excellent generalization performance of our method on compressed images while outperforming state-of-the-art techniques.

1. Introduction

With the significant improvement in computer performance and the rapid development of data mining and machine learning technologies, artificial intelligence is widely applied in many fields. However, despite enriching daily life, the development of technology has also brought some potential risks (Okey et al., 2022; Guo et al., 2023; Chi et al., 2022). Since its inception in 2018, Deepfake has advanced from amateur experimentation to a potentially malicious tool for facial feature manipulation. While Deep Learning (DL) advances, Deepfakes are also growing increasingly realistic. Besides, commercial applications such as ZAO, etc. are designed to be user-friendly such that the learning cost for individuals to use Deepfakes has dropped significantly. And this is followed by various malicious applications of Deepfake. During the Russian–Ukrainian conflict, a video featuring Ukrainian President Volodymyr Zelensky calling on Ukrainian soldiers to lay down their arms circulated widely. However, it was later confirmed that the video was a Deepfake and a mere rumor. This incident indicates that Deepfakes have been used in cognitive warfare. Additionally, Deepfake porn is one of malicious applications. According to a report by Wired magazine, every month thousands of people use DeepNude to create nude images of their friends and family, some of whom are under 18 years old. Therefore, it is urgent to design efficient and accurate methods to detect these Deepfakes.

Early approaches to Deepfake detection (Haliassos et al., 2021; Jung et al., 2020; Ciftci et al., 2020; Agarwal et al., 2020) are mostly based on handcrafted features such as heartbeat, blink, and lip shape, and use classifiers such as Support Vector Machine (SVM) and Random Forest (RF) for testing. However, handcrafted features primarily focus on sensitive facial information and may not be appropriate for all datasets, resulting in failures when the source of the data changes. To address these limitations, current research is heavily focused on Deep Learning methods (Gu et al., 2022; Sun et al., 2022; Dong et al., 2022).

Existing deep-feature-based methods typically perform well on high-quality images, but their performance deteriorates when testing on compressed images. This is because certain artifacts that could be used as detection cues in the fake image are weakened or removed during the compression process. Additionally, these models often extract features that are limited in scope, capturing either spatial or frequency domain information. This narrow focus can lead to overfitting, resulting in poor model generalization ability.

* Corresponding author.
E-mail address: songzr0518@163.com (Z. Song).

https://doi.org/10.1016/j.engappai.2024.108341
Received 17 November 2023; Received in revised form 3 March 2024; Accepted 25 March 2024
Available online 26 April 2024

Based on the above observations, we suggest that two factors should be considered when detecting compressed images: (1) restoring network attention to artifacts that may have been weakened during the compression process, and (2) making comprehensive use of information from both spatial and frequency domains. To address these challenges, we propose a Deepfake detection method that uses spatial-frequency fusion features and knowledge distillation.

Our approach begins with the introduction of a single branch called the Spatial-Frequency Fusion Branch (SFFB). This branch extracts spatial and frequency domain features from the input image and fuses them under attention guidance. In contrast to traditional knowledge distillation, our method employs the same structure for both the teacher and student models, with both models adopting the SFFB. Following the setup in ADD (Woo et al., 2022), we train the student model using paired input images of compressed and raw images while training the teacher model using only raw images. During the knowledge distillation process, the teacher model transfers both spatial and frequency domain features, as well as logits, to the student model, enhancing the network's attention to detect weakened artifacts.

It is worth noting that ADD also employed a knowledge distillation structure with paired inputs and achieved impressive results. However, unlike ADD, which requires paired data for training, our approach has the advantage of being able to achieve competitive results with a single input, making our model more flexible. In addition, we incorporate soft targets output from the teacher model to supervise the student model during knowledge distillation, enhancing the student model's ability to learn the teacher model comprehensively.

The contributions of this paper could be summarized as three-fold:

- We propose a Spatial-Frequency Fusion Branch (SFFB), the framework of which is simple and easy to implement. SFFB can achieve competitive results even when used independently.
- We apply knowledge distillation to deepfake detection. In the training process, we utilize spatial features, frequency domain features and logits for multi-knowledge transfer.
- Extensive visualizations and experiments demonstrate that our framework outperforms well-known baselines on the public datasets.

2. Related work

Over time, various techniques have been developed to detect Deepfakes, and the use of spatial and frequency domain features has proven to be a popular approach.

Spatial domain based Deepfake detection. This approach was prevalent in early works. In Afchar et al. (2018), Chollet (2017), Tan and Le (2019) and He et al. (2016), RGB images are used as input and deep features are extracted, leading to promising results on public Deepfake datasets. These models are now commonly used as backbone networks, contributing significantly to subsequent works. Recently, transformer (Vaswani et al., 2017) has become a hit in computer vision. Studies (Liang et al., 2023; Anas Raza et al., 2023; Ilyas et al., 2023) have shown that applying transformer in the field of deepfake detection can also achieve great performance. Wang et al. (2022), Liu et al. (2020) and Yu et al. (2019) extract artifact features like color cues, GAN fingerprints, and textures. In Shiohara and Yamasaki (2022) and Zhao et al. (2021), the authors propose methods to improve the testing accuracy of the detection model by designing the training data. Recently, disentanglement has gained significant attention in the field of Deepfake detection (Yan et al., 2023; Liang et al., 2022). By separating irrelevant information, the model's generalization ability is further improved. These methods perform well on raw or high-quality images. However, since they only focus on the information extracted from the spatial domain, their performance can be affected by high compression.

Frequency domain based Deepfake detection. In recent years, researchers have focused on solving the generalization problem and extracting frequency domain information. For example, Jia et al. (2021) trained a two-branch network based on Stationary Wavelet Decomposition (SWD) to extract inter-image and intra-image inconsistencies. Qian et al. (2020) proposed the Frequency in Face Forgery Network (F3-Net) that applies Discrete Cosine Transform (DCT) as the frequency-domain transformation. F3-Net uses complementary frequency-aware clues to achieve better performance in detecting low-quality forgeries. Frank et al. (2020) leveraged 2D-DCT to extract frequency domain information and efficiently detect GAN-generated images. These methods have shown that frequency domain information can provide complementary and useful clues for Deepfake detection. There is still further room for improvement in the performance of these methods because they only utilize information from single frequency domain.

Spatial-frequency fusion based Deepfake detection. To capture more comprehensive information, researchers have started to fuse spatial and frequency domain features. Liu et al. (2021) proposed a shallow network that combines spatial images and phase spectrum to capture up-sampling artifacts. Li et al. (2021) introduced a novel loss called the single-center loss to improve intra-class compaction and inter-class separability. They also designed an RGB-frequency fusion module, and the effectiveness of fusion has been demonstrated in ablation studies. Luo et al. (2021) fused high-frequency features extracted by SRM with spatial features, and their detection model achieved admirable results in cross-dataset evaluation. Gu et al. (2022) proposed a two-branch Deepfake detection network with enhancement learning for fine-grained feature extraction in the spatial and frequency domains. Liang et al. (2023) designed a two-stream framework which incorporates a spatial stream and a frequency stream, and a combination of coarse and fine classification was used to detect the Deepfakes. Shuai et al. (2023) proposed an innovative two-stream network and three functional modules to enlarge the potential regions from which the model extracted forgery evidence.

Knowledge Distillation. Knowledge distillation (KD) is a model compression technique that aims to derive small yet highly accurate networks from larger ones. It has recently been widely used in computer vision applications such as object detection (Zhang and Ma, 2021), semantic segmentation (Liu et al., 2019), and Deepfake detection (Woo et al., 2022; Kim et al., 2021; Xu et al., 2023). In our work, unlike traditional knowledge distillation, our goal is not model compression. We employ the same spatial-frequency fusion branch for both the teacher and student models, but the inputs of the two models are raw and compressed images, respectively. This setup allows the student model to focus on the weakened artifacts caused by compression.

Unlike methods that solely utilize features from one domain, our approach combines the strengths of both spatial and frequency domain, resulting in a more comprehensive and robust feature space. Compared with other spatial-frequency fusion based Deepfake detection methods, our method introduces knowledge distillation, extensively exploring the forged clues in uncompressed images while other models neglect the correlation between the original and compressed images. Multi-knowledge distillation enables our model to accurately locate forgery regions within compressed images while minimizing interference from non-forgery areas.

3. The proposed method

3.1. Overview

Aiming at solving the problems of previous methods in performance degradation with compressed Deepfake images, we propose a novel Deepfake detection method that integrates spatial and frequency domain information within a knowledge distillation framework. As illustrated in Fig. 1, knowledge distillation is applied as the overarching backbone of our framework. Both the teacher model and the student model employ the same SFFB structure, which extracts features from the spatial domain and frequency domain and fuses the features

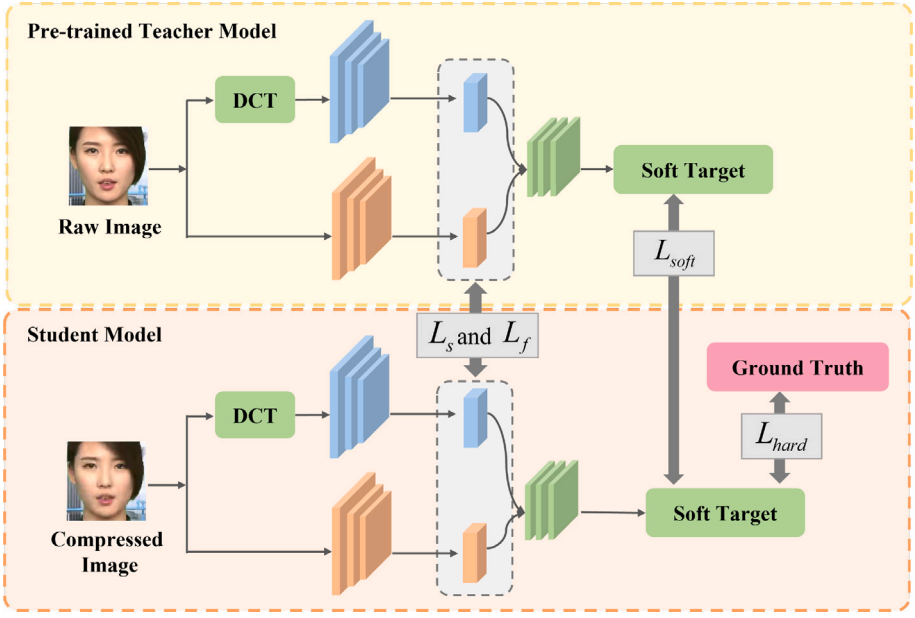


Fig. 1. The pipeline of our proposed method. Both the teacher model and the student model adopt the same SFFB structure, as shown in Fig. 2. The teacher model has been pretrained using the raw images. During the training phase, the teacher model takes the original image as input, while the student model takes the compressed image as input. In the knowledge distillation process, the feature maps of the spatial and frequency domains and soft target are transferred as knowledge to the student model.

through a fusion module. It is worth noting that the inputs of teacher model and student model are different, raw images for the teacher model, compressed images for the student model. In the knowledge distillation process, the feature maps of the spatial and frequency domains are transferred as knowledge to the student model. Furthermore, after going through the fusion module and convolution layers, logits are obtained, which are also transferred as knowledge.

3.2. Spatial-Frequency Fusion Branch

In order to obtain comprehensive features to improve the generalization ability of the network, we develop a Spatial-Frequency Fusion Branch (SFFB), as shown in Fig. 2. Our SFFB consists of four parts: data preprocessing module, feature extraction module, fusion module and classification module.

Data preprocessing. Overfitting is one of the causes of poor model generalization performance. To avoid overfitting, we first perform data augmentation on RGB images, including image flipping, grayscale, color jittering and so on. The images after data augmentation are used as input to the spatial stream directly. We denote the RGB image as $I_s^{s/t}$ (superscript: s for student, t for teacher). For frequency information mining, 2D-DCT is applied to the image. Then the frequency domain image is going through three filters for different bands (low frequency, medium frequency and high frequency). Finally, the image is reconstituted by the DCT inverse transformation. The process of frequency domain conversion can be formulated as:

$$I_{fi}^{s/t} = f_i\left(DCT\left(I_s^{s/t}\right)\right), \tag{1}$$

$$I_f^{s/t} = concat\left(DCT^{-1}\left(I_{fi}^{s/t}\right)\right), \tag{2}$$

where $f_i(\cdot)$ denotes the filter, $I_{fi}^{s/t}$ denotes the single image after the filter, $concat(\cdot)$ denotes that concatenation along the channel direction, and $I_f^{s/t}$ is the final frequency domain input.

Feature extraction module. Then $I_f^{s/t}$ and $I_s^{s/t}$ are fed into the feature extraction module. After that, frequency feature map $M_f^{s/t}$ and spatial map $M_s^{s/t}$ are obtained. We choose the entry and middle flow of Xception (Chollet, 2017) as the backbone network for feature extraction. It is worth noting that the weights of the spatial stream and frequency stream are not shared.

Fusion module. Then the obtained spatial and frequency domain feature maps are going through the fusion module. The fusion module is inspired by SKAttention (Li et al., 2019), which consists of three parts: split, fuse and select. As illustrated in Fig. 3, different from SKAttention, the "split" part is replaced with the obtained spatial and frequency maps in our fusion module. Specifically, we first apply element-wise summation to the two feature maps and obtain the mixed feature map $M_m^{s/t}$:

$$M_m^{s/t} = M_f^{s/t} + M_s^{s/t}, \tag{3}$$

the purpose of which is to mix information from different domains. Furthermore, because both streams have the same structure except the data preprocessing module, for each feature map of different domains, the features at the same position are correlated, which means that the element-wise summation will not disturb the original feature arrangement. Then do the channel reduction, which consists of two steps. First step is global average pooling (GAP) (Lin et al., 2013), and this step can be formulated as:

$$A_{r1}^{s/t} = concat\left(\frac{1}{|R|}\sum_{(p,q)\in R} x_{kpq}^{s/t}\right), k = 1, 2, \ldots, n, \tag{4}$$

where $x_{kpq}^{s/t}$ denotes the element at (p,q) in the kth feature map of $M_m^{s/t}$, n denotes the number of feature maps for $M_m^{s/t}$, $|R|$ denotes the elements number of the feature map, and $A_{r1}^{s/t}$ denotes the output after GAP. Secondly, a fully connected layer is followed:

$$A_{r2}^{s/t} = fc\left(A_{r1}^{s/t}\right), \tag{5}$$

where $fc(\cdot)$ denotes the fully connected operation. This step allows information between different channels to interact. To match the size of input feature maps, another two fully connected layer ($fc_1(\cdot)$ and $fc_2(\cdot)$) are applied to A_{r2} and the corresponding features are element-wise multiplied with the spatial feature map and frequency map respectively, $A_f^{s/t}$ and $A_s^{s/t}$ obtained as:

$$A_f^{s/t} = fc_1\left(A_{r2}^{s/t}\right) \otimes M_f^{s/t} + M_f^{s/t}, \tag{6}$$

$$A_s^{s/t} = fc_2\left(A_{r2}^{s/t}\right) \otimes M_s^{s/t} + M_s^{s/t}, \tag{7}$$

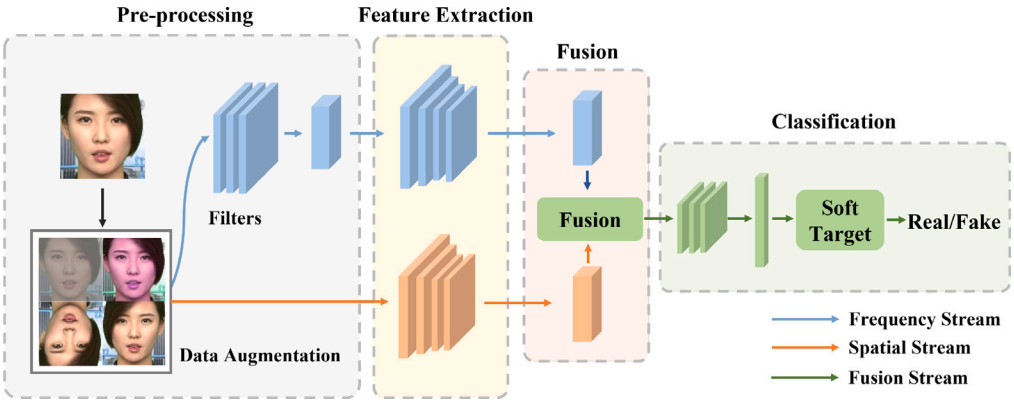


Fig. 2. Illustration of our Spatial-Frequency Fusion Branch (SFFB). SFFB consists of four parts: data preprocessing module, feature extraction module, fusion module and classification module. The spatial and frequency domain features from fusion module, as well as the soft target from the classification module, are transferred as knowledge.

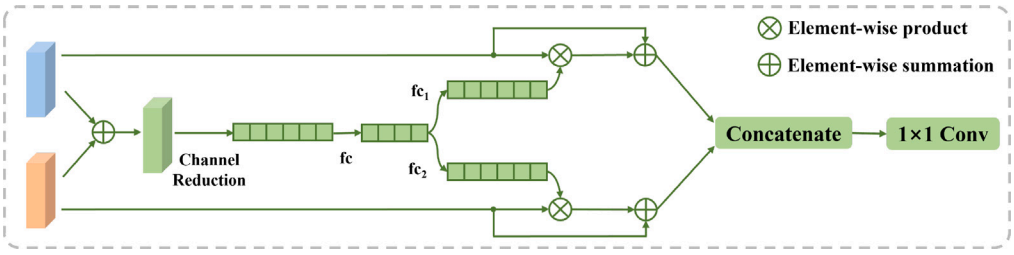


Fig. 3. Illustration of the fusion module. The fusion module consists of fusion, selection and concatenation.

Different from SKAttention, we add element-wise summation of the element-wise product results with corresponding original feature map to enhance the effect of the fusion. Then, we concatenate the features from the two domains in the direction of channel and a 1×1 convolution layer is applied to fuse the features, and the fused feature $M_{m_1}^{s/t}$ for classification is obtained as:

$$M_{m_1}^{s/t} = conv\left(concat\left(A_f^{s/t}, A_s^{s/t}\right)\right), \tag{8}$$

where $conv(\cdot)$ denotes the 1×1 convolution.

Classification. Finally, we choose the exit flow of Xception for classification. The SFFB branch can be trained independently for Deepfake detection, and we present its performance in Sections 4.2 and 4.3.

3.3. Knowledge distillation for SFFB

The framework of our knowledge distillation is shown in Fig. 1. Compared to traditional knowledge distillation, our model differs in three ways: (1) input, (2) the size of the models, and (3) knowledge. First, the teacher model is pre-trained on raw images. In the training stage of the student model, the input of the teacher model is raw image and the model weights are fixed. For the student model, input is the corresponding compressed image. However, in traditional knowledge distillation, the input of the teacher model and the student model is the same. Further, different from traditional knowledge distillation which aims to perform model compression, in our work, both the teacher model and the student model adopt the same SFFB structure, because we focus on the knowledge transfer between raw images and compressed images. Knowledge is usually divided into three types: response-based knowledge, relation-based knowledge and feature-based knowledge. We not only make the student model imitate the final predictions of the teacher model, but also let the student model learn the feature representations of the teacher model, so we adopt both response-based knowledge and feature-based knowledge simultaneously.

We select the spatial and frequency domain feature maps before the fusion module as feature-based knowledge which is supervised by Mean Square Error (MSE). The loss here can be formulated as:

$$L_f = \frac{1}{|R|}\sum_{i=1}^{|R|}\left(M_{fi}^s - M_{fi}^t\right)^2, \tag{9}$$

$$L_s = \frac{1}{|R|}\sum_{i=1}^{|R|}\left(M_{si}^s - M_{si}^t\right)^2. \tag{10}$$

The KD loss for response-based knowledge can be formulated as:

$$L_{soft} = -\sum_{i}^{N} x_i^T \log\left(y_i^T\right), \tag{11}$$

where x_i^T refers to the value of the softmax output of the teacher model on class i when temperature parameter is T, and y_i^T refers to the value of the softmax output of the student model. x_i^T and y_i^T are obtained as:

$$x_i^T = \frac{e^{v_i/T}}{\sum_k^N e^{v_k/T}}, \tag{12}$$

$$y_i^T = \frac{e^{z_i/T}}{\sum_k^N e^{z_k/T}}, \tag{13}$$

where v_i is the logit of the teacher model, z_i is the logit of the student model, and N refers to the total number of labels. Additionally, we adopt the widely-used Cross-Entropy loss for ground truth supervision:

$$L_{hard} = -\sum_{i}^{N} c_i \log\left(y_i^1\right), \tag{14}$$

where c_i refers to the ground truth label on class i, $c_i \in \{0, 1\}$, and y_i^1 denotes the prediction of the student model. The purpose of hard loss is to reduce the possibility of errors being propagated to the student model, as the teacher model also has a certain error rate, and using ground truth can effectively lower the chance of errors being passed on. The total loss is written as:

$$L_{total} = \lambda_1\left(L_f + L_s\right) + \lambda_2 L_{soft} + \lambda_3 L_{hard}, \tag{15}$$

Table 1
The ACC results of our proposed method on FF++ dataset (compressed).

Methods	c23				c40			
	DF	FS	F2F	NT	DF	FS	F2F	NT
Steg.Features (Fridrich and Kodovsky, 2012)	77.12	79.51	74.68	76.94	65.58	60.58	57.55	60.69
Cozzolino et al. (2018)	81.07	82.25	79.26	80.38	70.17	62.22	60.90	63.24
F3-Net (Qian et al., 2020)	96.26	97.85	95.52	77.91	93.06	92.49	81.48	61.95
MesoNet (Afchar et al., 2018)	89.77	95.50	94.25	78.70	77.68	79.92	83.65	77.74
Xception (Chollet, 2017)	95.15	95.96	97.07	87.99	83.70	83.17	**87.21**	87.90
ResNet50 (He et al., 2016)	96.34	92.46	95.60	86.25	92.89	88.91	83.94	60.27
EfficientNetV2 (Tan and Le, 2021)	88.44	90.22	90.00	85.49	79.62	71.74	71.88	84.11
ConvNeXt (Liu et al., 2022)	91.99	90.97	92.01	88.25	82.21	79.70	72.72	87.70
ADD (Woo et al., 2022)	98.67	97.85	96.82	88.48	**95.50**	**92.49**	85.42	68.53
Ours-SFFB	97.46	97.65	94.33	95.31	88.90	82.73	83.45	98.13
Ours-SFFB&KD	**98.87**	**98.57**	**98.58**	**96.84**	92.00	89.22	85.63	**98.36**

Table 2
Experimental results of SFFB on FF++ dataset (raw) and Celeb-DF dataset.

Methods	DF		FS		F2F		NT		Celeb-DF	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
D-CNN (Rahmouni et al., 2017)	98.03	–	98.94	–	98.96	–	96.06	–	–	–
MesoNet (Afchar et al., 2018)	96.37	98.82	98.17	99.48	97.95	98.90	93.30	96.87	85.30	90.61
Xception (Chollet, 2017)	98.31	98.98	97.10	98.04	97.75	99.15	96.45	97.49	99.96	**100.00**
Ours-SFFB	**99.13**	**99.93**	**99.20**	**99.75**	**99.16**	**99.77**	**99.09**	**99.93**	**99.98**	**100.00**

where λ_1, λ_2 and λ_3 act as hyperparameters, controlling the trade-off between different loss functions.

4. Experiments

In this section, we first present the overall experimental setup, and then present extensive experimental results and visualization to demonstrate the effectiveness of our method.

4.1. Experimental setup

Dataset. We adopt two widely-used public datasets in our experiments, FaceForensics++ (FF++) (Rossler et al., 2019) and Celeb-Deepfake (Celeb-DF) (Li et al., 2020). FaceForensics++ is a large-scale dataset containing over 100,000 synthesized videos, consists of four types of face manipulations: Deepfake (DF), Face2Face (F2F), FaceSwap (FS) and NeuralTexture (NT). The videos in FF++ dataset are compressed into two versions: light compression (c23) and heavy compression (c40), obtained by the H.264 codec with a constant rate quantization parameter of 23, and 40, respectively. Celeb-DF consists of 408 original videos downloaded from YouTube and 795 fake videos. For FF++ dataset, we randomly selected 720 videos for training, 140 videos for validation and 140 videos for test. We divide Celeb-DF dataset into training set, validation set and test set in the ratio of $6:1:1$.

Evaluation metrics. We apply Accuracy score (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC) as our evaluation metrics.

Implementation detail. First, we pretrain the teacher model, that is, training a single-branch SFFB on raw images. Then, we train the student model where the teacher model takes raw images as input, and the student model takes compressed images as input. In addition, to ensure the guidance of the teacher model on the student model, the input raw images and compressed images are paired. It is worth noting that the weights of the teacher model remain unchanged throughout the entire training process, while the student model has no initial weights.

We implemented the proposed method with PyTorch. We sample 50 frames per video and resize the frames to 256×256. Xception (Chollet, 2017) pretrained on ImageNet is employed as the SFFB backbone. To be precise, the entry flow and middle flow of Xception are used for feature extraction, and the exit flow is applied after feature fusion for classification. We employ Adam optimizer with the learning rate of 2×10^{-4}, and the batch size of 32. The hyper-parameters in L_{total} are $\lambda_1 = 2$, $\lambda_2 = 100$, $\lambda_3 = 1$. Temperature parameter T is set to 1.

4.2. Experimental results

We firstly test our model on different qualities of FF++ dataset. Specifically, we pretrain the teacher model using raw images, and train our student model respectively in each compression level (c23 and c40). We compare our method with other Deepfake detection methods, including ADD which also employs knowledge distillation. The results are presented in Table 1. Fridrich and Kodovsky (2012) and Cozzolino et al. (2018) are traditional deepfake detection methods, and their drawback is that the testing performance of the model is easily influenced by the quality of the images. Models such as Xception, ResNet50 and EfficientNetV2 often serve as CNN baselines and does not incorporate any augmentation or frequency information. Its performance drops dramatically when testing with compressed images. We can see from the Table 1 that, in most cases, our method outperforms others in ACC. Our model achieved an average ACC of 98.22% and 91.30% on FF++ c23 and c40 version respectively. Especially in the c23 version, compared to the baseline method ADD, the average ACC is improved by 2.77%, which means our method achieves state-of-the-art results. This demonstrates the feasibility of multi-knowledge distillation and spatial-frequency fusion.

NeuralTexture is a powerful deepfake generation technology that can extract textures from one image and apply them to another image using neural networks. This process guarantees the high resolution and highly detailed texture of output fake images. Because the original colors and details are retained, detecting fake images generated by NeuralTexture efficiently remains a significant challenge. But in our work, even on the most challenging NT c40 version, we achieve a 98.36% ACC score, nearly 30% ACC improvement compared to ADD. We can observe from Table 1 that even using SFFB alone can achieve 98.13% accuracy on NT c40 version, which proves that the features extracted by our Spatial-Frequency Fusion Branch are comprehensive and efficient.

To demonstrate the flexibility of our method, we test SFFB on the c0 version of FF++ dataset and Celeb-DF dataset. As shown in Table 2, we can observe that our SFFB can still achieve competitive results even without KD, and achieve the state-of-the-art with near-perfect performance on FF++ dataset(raw) and Celeb-DF dataset. And

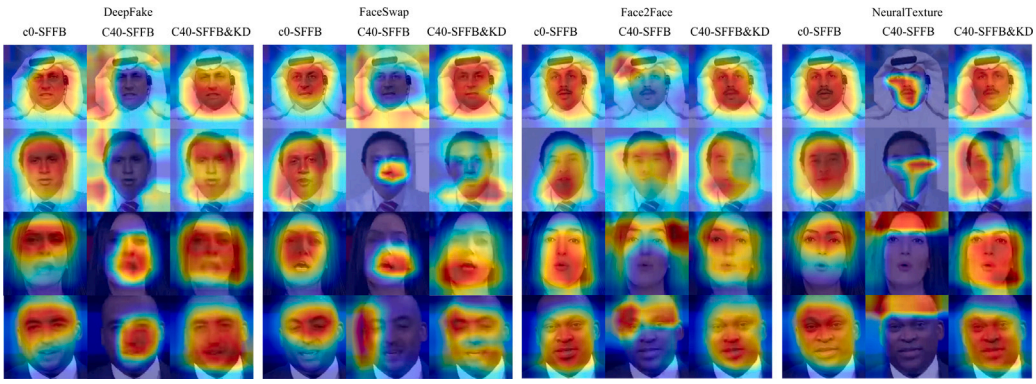


Fig. 4. Visualizations of our model.

Table 3
The performance of different fusion modules on the F2F c40 version.

Fusion module	ACC
Summation	83.00
Concatenation	81.65
Chunked attention	71.74
Ours	**83.45**

Table 4
The performance of different knowledge types on the F2F c40 version.

Knowledge types	ACC
Summation of $M_s^{s/t}$ and $M_f^{s/t}$	83.79
Fusion of $M_s^{s/t}$ and $M_f^{s/t}$	84.14
$M_s^{s/t}$ and $M_f^{s/t}$	83.38
$M_s^{s/t}$, $M_f^{s/t}$ and logits	**85.63**

Table 5
The performance of different backbones (ACC).

Methods	DF		FS		F2F		NT	
	c23	c40	c23	c40	c23	c40	c23	c40
ResNet50	95.51	86.01	97.48	79.96	97.29	76.27	94.54	93.30
Ours-ResNet50	96.82	88.74	98.06	86.28	98.34	83.30	95.21	95.72
ResNet34	96.04	86.41	97.24	79.97	97.15	79.70	92.68	95.35
Ours-ResNet34	96.39	86.43	97.56	84.35	97.36	81.54	93.36	95.17

it is worth mentioning that our method achieves 100% AUC on Celeb dataset. Seeing the results from Tables 1 and 2, our SFFB achieves average ACC of 99.15%, 96.19%, and 88.30% on the compression levels c0, c23, and c40 of FF++ dataset respectively. Although the results are worse than SFFB&KD, they are still competitive compared with other works.

Although we have achieved remarkable results on compressed images, there exist limitations of our model. In spite of the great performance on the challenging NT c40 dataset, the test results of our model on the DF and FS datasets are lower than those of the baseline model ADD. Our results on the F2F dataset are better than ADD but lower than the backbone network Xception. The probable cause is overfitting.

4.3. Ablation study

To demonstrate the effects of each module in our work, we perform ablation study on FF++ dataset.

Effect of knowledge distillation. Although SFFB can achieve competitive results when used alone, there is still room for improvement in performance. As shown in Table 1, After applying knowledge distillation, the model achieved a maximum improvement of 6.49% in ACC. The improvement in model performance after applying knowledge distillation is due to the fact that the distilled information provides additional guidance to the student model during training, allowing it to learn more effectively and generalize better to compressed data.

Effect of fusion module. We conduct this experiment using SFFB independently to evaluate the effectiveness of different fusion module structures. As shown in Table 3, summation means we adopt element-wise summation directly to the $M_s^{s/t}$ and $M_f^{s/t}$, and concatenation means we concatenate the feature maps of the two domains along the channel dimension and use a 1×1 convolution to do the dimensionality reduction. Chunked attention contains three steps: (1) divide the frequency domain and spatial domain feature maps into an equal number of blocks, (2) traverse each block, and fuse its corresponding frequency domain and spatial domain blocks following the procedure shown in Fig. 3 to obtain a fused block, and (3) concatenate all the fused blocks according to their positions in the original feature map to obtain the fused feature map. We can observe that our approach has the best performance due to the incorporation of attention guidance in our fusion module, which enables the student model to effectively learn the useful information from the teacher model. The reason why chunked attention does not work well may be the global information loss caused by chunk.

Effect of multi-knowledge. In order to explore the influence of different knowledge types during knowledge distillation, we conduct experiments on F2F dataset of c40 version. As shown in Table 4, the best performance is achieved when all $M_s^{s/t}$, $M_f^{s/t}$ and logits are transferred simultaneously because the student model can learn from the teacher model more comprehensively.

Effect of different backbones. To further demonstrate the effectiveness of our method, we conducted experiments using different backbones and the results are shown in Table 5. During the experiments, we ensure the consistency of transferred feature sizes among different backbone networks. Table 5 shows that in most cases, our method improved the accuracy compared to the corresponding baseline with SFFB.

Visualizations. We adopt grad-cam to visualize the features of the samples from the FF++ dataset. As shown in Fig. 4, we can observe that when the image is raw, SFFB can focus on the tampered facial area, which is the reason why the SFFB branch can achieve great testing performance on raw images. However, when the image quality drops to heavy compression (c40), the attention of SFFB is dispersed to areas such as hair and background or overly focused on a small region, resulting in a significant degradation in performance. When we use SFFB and multi-knowledge distillation simultaneously, guided by the teacher model, the network's attention is restored to the tampered facial area, and the problem of small attention regions has also been alleviated, which ensures the generalization ability of the model. However, we must admit that our method's attention region is still not

precise enough in some cases, and a small portion of the network's attention is still scattered in the background area (see the second row in Fig. 4).

5. Conclusion

In this paper, we propose a spatial-frequency fusion based on knowledge distillation for Deepfake detection. Comprehensive extraction of spatial-frequency features ensures great generalization performance of the model on compressed images. In addition, the incorporation of knowledge distillation increases the student model's attention to weakened spatial-frequency features. The experimental results demonstrate the great performance on compressed data. But our model still has some limitations, for example, there is room for improvement in the experimental results on heavy compressed data. In the future, we will explore the more effective utilization of knowledge distillation in the field of Deepfake detection.

CRediT authorship contribution statement

Bo Wang: Methodology, Conceptualization. **Xiaohan Wu:** Writing – original draft, Methodology. **Fei Wang:** Visualization, Validation, Software. **Yushu Zhang:** Resources, Project administration. **Fei Wei:** Writing – review & editing, Resources, Funding acquisition. **Zengren Song:** Writing – review & editing, Validation, Supervision, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Funding

This work is supported by the National Natural Science Foundation of China (No. 62106037, No. 62076052); the Science and Technology Innovation Foundation of Dalian, China (No. 2021JJ12GX018); the Application Fundamental Research Project of Liaoning Province, China (2022JH2/101300262); and the Major Program of the National Social Science Foundation of China (No. 19ZDA127).

References

Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., 2018. Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security. WIFS, IEEE, pp. 1–7.

Agarwal, S., Farid, H., El-Gaaly, T., Lim, S.-N., 2020. Detecting deep-fake videos from appearance and behavior. In: 2020 IEEE International Workshop on Information Forensics and Security. WIFS, IEEE, pp. 1–6.

Anas Raza, M., Mahmood Malik, K., Ul Haq, I., 2023. HolisticDFD: Infusing spatiotemporal transformer embeddings for deepfake detection. Inform. Sci. 645, 119352. http://dx.doi.org/10.1016/j.ins.2023.119352, URL https://www.sciencedirect.com/science/article/pii/S0020025523009374.

Chi, H.R., Wu, C.K., Huang, N.-F., Tsang, K.-F., Radwan, A., 2022. A survey of network automation for industrial internet-of-things toward industry 5.0. IEEE Trans. Ind. Inform. 19 (2), 2065–2077.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Ciftci, U.A., Demir, I., Yin, L., 2020. Fakecatcher: Detection of synthetic portrait videos using biological signals. IEEE Trans. Pattern Anal. Mach. Intell..

Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L., 2018. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510.

Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Yu, N., Chen, D., Wen, F., Guo, B., 2022. Protecting celebrities with identity consistency transformer. arXiv preprint arXiv:2203.01318.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T., 2020. Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning. PMLR, pp. 3247–3258.

Fridrich, J., Kodovsky, J., 2012. Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. 7 (3), 868–882.

Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., Yi, R., 2022. Exploiting fine-grained face forgery clues via progressive enhancement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 735–743.

Guo, R., Liu, H., Liu, D., 2023. When deep learning-based soft sensors encounter reliability challenges: A practical knowledge-guided adversarial attack and its defense. IEEE Trans. Ind. Inform..

Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M., 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5039–5049.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Ilyas, H., Javed, A., Malik, K.M., 2023. AvFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. Appl. Soft Comput. 136, 110124. http://dx.doi.org/10.1016/j.asoc.2023.110124, URL https://www.sciencedirect.com/science/article/pii/S1568494623001424.

Jia, G., Zheng, M., Hu, C., Ma, X., Xu, Y., Liu, L., Deng, Y., He, R., 2021. Inconsistency-aware wavelet dual-branch network for face forgery detection. IEEE Trans. Biom. Behav. Identity Sci. 3 (3), 308–319.

Jung, T., Kim, S., Kim, K., 2020. Deepvision: Deepfakes detection using human eye blinking pattern. IEEE Access 8, 83144–83154.

Kim, M., Tariq, S., Woo, S.S., 2021. Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1001–1012.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 510–519.

Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y., 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6458–6467.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207–3216.

Liang, J., Shi, H., Deng, W., 2022. Exploring disentangled content information for face forgery detection. In: European Conference on Computer Vision. Springer, pp. 128–145.

Liang, Y., Wang, M., Jin, Y., Pan, S., Liu, Y., 2023. Hierarchical supervisions with two-stream network for Deepfake detection. Pattern Recognit. Lett. 172, 121–127. http://dx.doi.org/10.1016/j.patrec.2023.05.029, URL https://www.sciencedirect.com/science/article/pii/S0167865523001678.

Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J., 2019. Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2604–2613.

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N., 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 772–781.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Liu, Z., Qi, X., Torr, P.H., 2020. Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8060–8069.

Luo, Y., Zhang, Y., Yan, J., Liu, W., 2021. Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16317–16326.

Okey, O.D., Maidin, S.S., Adasme, P., Lopes Rosa, R., Saadi, M., Carrillo Melgarejo, D., Zegarra Rodríguez, D., 2022. BoostedEnML: Efficient technique for detecting cyberattacks in IoT systems using boosted ensemble machine learning. Sensors 22 (19), 7409.

Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J., 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII. Springer, pp. 86–103.

Rahmouni, N., Nozick, V., Yamagishi, J., Echizen, I., 2017. Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE Workshop on Information Forensics and Security. WIFS, IEEE, pp. 1–6.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11.

Shiohara, K., Yamasaki, T., 2022. Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18720–18729.

Shuai, C., Zhong, J., Wu, S., Lin, F., Wang, Z., Ba, Z., Liu, Z., Cavallaro, L., Ren, K., 2023. Locate and verify: A two-stream network for improved deepfake detection. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7131–7142.

Sun, K., Yao, T., Chen, S., Ding, S., Li, J., Ji, R., 2022. Dual contrastive learning for general face forgery detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 2316–2324.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tan, M., Le, Q., 2021. Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning. PMLR, pp. 10096–10106.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. ArXiv.

Wang, B., Li, Y., Wu, X., Ma, Y., Song, Z., Wu, M., 2022. Face forgery detection based on the improved siamese network. Secur. Commun. Netw. 2022, 1–13.

Woo, S., et al., 2022. ADD: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 122–130.

Xu, X., Tang, S., Zhu, M., He, P., Li, S., Cao, Y., 2023. A novel model compression method based on joint distillation for deepfake video detection. J. King Saud Univ. - Comput. Inf. Sci. 35 (9), 101792. http://dx.doi.org/10.1016/j.jksuci.2023.101792, URL https://www.sciencedirect.com/science/article/pii/S1319157823003464.

Yan, Z., Zhang, Y., Fan, Y., Wu, B., 2023. UCF: Uncovering common features for generalizable deepfake detection. arXiv preprint arXiv:2304.13949.

Yu, N., Davis, L.S., Fritz, M., 2019. Attributing fake images to gans: Learning and analyzing gan fingerprints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7556–7566.

Zhang, L., Ma, K., 2021. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In: International Conference on Learning Representations.

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W., 2021. Learning self-consistency for deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15023–15033.