EXPLORING INTERACTIONS BETWEEN MODALITIES FOR DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As face forgery techniques have become more mature, the proliferation of deepfakes may threat the human society security. Although existing deepfake detection methods achieve a good performance for in-dataset evaluation, it still remains to be improved in the generalization abiltiy, where the representation of the imperceptible artifacts plays a significant role. In this paper, we propose an Interactive Two-Stream Network (ITSNet) to explore the discriminant inconsistency representation from the perspective of cross-modality. Specially, the patch-wise Decomposable Discrete Cosine Transform (DDCT) is adopted to extract fine-grained high-frequency clues and information from different modalities are communitcated with each other via a designed interaction module. To perceive the temporal inconsistency, we first develop a Short-term Embedding Module (SEM) to refine subtle local inconsistency representation between adjacent frames, and then a Long-term Embedding Module (LEM) is designed to further refine the erratic temporal inconsistency representation from the long-range perspective. Extensive experimental results conducted on three public datasets show that ITSNet outperforms the state-of-the-art methods both in terms of in-dataset and cross-dataset evaluations.

1 INTRODUCTION

Facial forgery methods and deepfake detection methods have made great progress due to mutual competition. For the forgery methods, the advent of deep generative models Korshunova et al. (2017); Bao et al. (2018); Natsume et al. (2018); Li et al. (2020a) and face manipulated methods (such as Deepfakes git (a) and Face2Face Thies et al. (2016a)) allow arbitrary tampering with other people's privacy. Therefore, it is crucial to develop discriminative deepfake detection methods to indicate the authenicity of images or videos, especially for the latest forgery methods.

In the early stages, hand-crafted forgery features Li et al. (2018); Yang (2019) as well as data-driven forgery features once dominated when dealing with rough fake faces. Although these methods Afchar et al. (2018); Hsu et al. (2018); Li et al.; Zhang et al. (2019) have achieved great success in particular datasets, insufficient attention has been paid to the generalization ability of unseen forgery methods.

Recent methods try to encode the common forgery trace caused by various forgery methods, such as temporal inconsistency features and frequency artificial features. For the former, since most forgery videos are manipulated on the individual frame which may bring temporal clues for the deepfake detection task. Some detection methods Güera & Delp (2018); Sabir et al. (2019); Agarwal et al. (2020); Zheng et al. (2021) pay attention to extracting discriminate temporal features from the consecutive video frames through 3DCNN-based or RNN-based methods. However, without being specifically designed, the limited temporal receptive field may lead to high computational cost and the faintish incoherent perception ability among long-distant frames. As for the frequency domain features, most methods Qian et al. (2020); Frank et al. (2020); Gu et al. (2021) pay attention to utilizing the frequency information to eliminate the irrelevant features of RGB image. While the long-range temporal inconsistency of the frequency features has not yet been fully exploited.

Different from these methods, we construct an interactive two-stream network (ITSNet) by embedding the high-frequency distribution of the video from both the frequency modality and the RGB modality. Specifically, we design a Decomposable Discrete Cosine Transform (DDCT) to generate

fine-grained frequency features and filter out irrelevant low-frequency components, which provides more significant frequency representation. Furthermore, a short-term embedding module (SEM) is developed to encode the local temporal information in the early stage. In the later stage of our proposed network, a long-term embedding module (LEM) is designed to summarize the global temporal feature of the video and the perception field in the time dimension is further broadened. At the end of each stage, a cross-modality interaction is conducted and the representation for each modality can be enhanced by sensing the inconsistency across modalities. In summary, our contributions are three-fold:

- We propose an interactive two-stream network (ITSNet) to detect face forgeries in videos. Two modalities including the high-frequency component of the video and the RGB information are adopted and they are communicated through a designed interaction module.
- In the frequency stream, we design a patch-wise Decomposable Discrete Cosine Transform (DDCT), which can represent effective high-frequency components to distinguish face forgery videos from real faces. For each stream, long-term inconsistency and short-term inconsistency are encoded by a short-term embedding module and a long-term embedding module.
- We conduct extensive experiments on FF++, CD2, and WildDeepfake against numerous state-of-the-arts and the experimental results demonstrate the superior detection performance and generalization ability of our method on face forgery video detection.

2 Related Work

In this section, we discuss previous related works in this field. In Section 2.1, we briefly introduce previous works on face forgery detection methods for videos. We specifically review related works that also employ frequency information and research on high-frequency components for facial forgery detection in Section 2.2.

2.1 Video-based forgery detection

With the rapid development of face forgery technology, the detection of forged face video has been extensively studied in recent years. Early works Yang (2019); Afchar et al. (2018); Chollet (2017) utilize deep 2D convolution neural network (CNN) to extract task-oriented spatial features for forged image detection. Although these vanilla CNN-based methods are simple to implement, it is not suitable for more realistic and unknown source counterfeit.

Later on, researchers begin to capture the temporal inconsistency from the video clip. Among them, a branch of works taking the explicit physiological signals Li et al. (2018); Agarwal et al. (2020); Haliassos et al. (2021); Sabir et al. (2019) as the basis to detect the temporal inconsistency in the forged frame sequential. For example, Li *et al.* Li et al. (2018) determine the authenity of videos by judging the eye blinking behavior in the video. Haliassos *et al.* Haliassos et al. (2021) detect inconsistencies of mouth movements by leveraging rich representations learned from the lip-reading task. Sabir *et al.* Sabir et al. (2019) concentrate on the discrepancy of landmark trajectory between real and fake videos with bidirectional-recurrent-denset. Such approaches can learn to detect temporal artifacts from additional annotations, while it is difficult to achieve satisfactory performance when dealing with occlusions and manipulations by unseen face forgery methods.

Another branch of works Güera & Delp (2018); Zheng et al. (2021) implicitly learn the temporal inconsistency between frame sequences in the RGB space. For instance, Güera *et al.* Güera & Delp (2018) utilize a RNN network to capture temporal inconsistencies between adjacent frames. By modifying the 3D convolution kernel into variants that are robust in time and shrink in space, Zheng *et al.* Zheng et al. (2021) make the 3DCNN more focused on coherence in the temporal dimension. However, a simple variation on the convolution kernel along the time dimension is still constrained by the receptive limitation of 3D convolution, which leads to the incomplete exploitation of long-term dependency. In contrast, the proposed ITSNet encodes both the long-range and short-term temporal inconsistency through cross-modality communication.

2.2 FREQUENCY-BASED FORGERY DETECTION

Besides the RGB domain, the frequency clues can also bring informative features for deepfake detection. Frequency reflects the variation of image brightness and can be further decomposed into different levels of components according to the local gradient. For the forgery detection task, Frank *et al.* Frank et al. (2020) observe that forged images generated by Generative Adversarial Networks (GAN) show particular artifacts in the frequency domain in the essential up-sampling operation. Durall *et al.* Durall et al. (2019) further prove that the discrepancy between real and fake faces is more evident in the high-frequency component than in the low-frequency component.

Apart from the vanilla frequency transform methods, such as Discrete Fourier Transform (DCT), a number of variants are proposed to fully exploit the underlying frequency artifacts in deepfake detection task. For example, Zhang *et al.* Zhang et al. (2019) extract frequency features from the residuals generated by median filtering of the image to detect forgery faces. Gu *et al.* Gu et al. (2021) view DCT in a fine-grained perspective by a vanilla sliding window. Qian *et al.* Qian et al. (2020) set certain constraints for the learnable filter to adaptively separate the frequency information. Although these frequency-based methods raise the bar of forged image detection, there is still a large room to improve the subtle frequency feature extraction and inconsistency processing in the temporal dimension. In our proposed ITSNet, we design a novel Decomposable DCT to extract fine-grained frequency information and a subsequent embedding module to represent the frequency inconsistency among the long-range sequential video frames.

3 INTERACTIVE TWO-STREAM NETWORK

This section starts with the network architecture of our proposed interactive two-stream model (ITSNet) in Sec. 3.1. We delinate the patch-wise decomposable DCT transform in Sec. 3.2. The short-term embedding module (SEM) and the long-term embedding module (LEM) are described in Sec. 3.3 and 3.4.

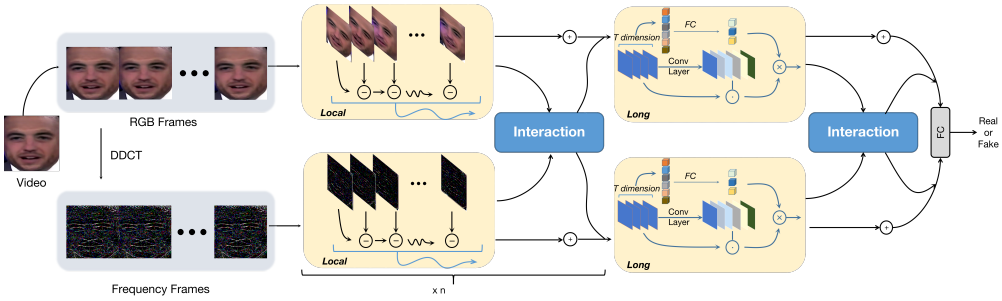

Figure 1: The architecture of the proposed ITSNet.

3.1 NETWORK ARCHITECTURE

The artificial traces of forgery face videos exist in both frequency and the RGB domain. To fully capture the traces, we propose an ITSNet containing two interactive branches to extract both frequency and textural temporal inconsistency, as described in Fig1. Firstly, the DDCT extracts fine-grained high-frequency components by performing the patch-level DCT convolution operation on the input RGB frames. Then the RGB frames with the extracted high-frequency frames are sent to the subsequent CNN-based parallel branch to encode long-range temporal inconsistency from the frequency perspective as well as the semantic view, respectively. Specifically, these two branches is constructed by a novel SEM in the early stage and a LEM in the late stage, which embeds the subtle temporal inconsistency from local to global. SEM narrows down the temporal receptive field and models the subtle motion information between adjacent frames. The LEM further expands the temporal perceptive field by assembling the local temporal inconsistency extracted by the SEM.

Given the feature map extracted from the RGB branch and Frequency branch, how to eliminate the misalignment between two modalities is indispensable. Inspired by Chen et al. (2021), we

consecutively utilize a spatial attention-based mechanism to promote the interaction between two branches and refine the subtle artifact perception capability of each individual modality. Denote the feature maps generated by the RGB branch and the frequency branch as $X^{RGB} \in R^{t\times c\times h\times w}$ and $X^{fre} \in R^{t\times c\times h\times w}$. We first concatenate the RGB and frequency feature map in the channel dimension, and they are comprehensively communicated by a 3×3 convolution layer to generate the cross-modality representation $X^{modality} \in R^{2\times H\times W}$, as formulated in Eq. 1.

$$X^{modality} = ReLU(BN(Conv_{3\times 3}(Concat(X^{RGB}, X^{freq})))) \tag{1}$$

After obtaining the cross-modality representation $X^{modality}$, we split it on the channel dimension and perform a channel-wise selection in each branch, as described in Eq. 2 and Eq. 3.

$$X^{RGB1} = X^{RGB} \odot split(X^{modality})_0 + X^{RGB} \tag{2}$$

$$X^{freq1} = X^{freq} \odot split(X^{modality})_1 + X^{freq} \tag{3}$$

Finally, a fully connected (FC) layer yields the forgery probability according to the final cross-modality representation, which is acquired by concatenating the long-range temporal inconsistency embedding and textural inconsistency embedding extracted by the LEM in each branch.

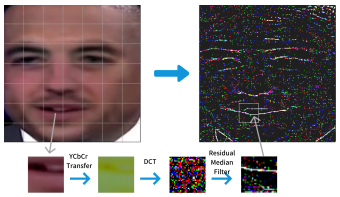

Figure 2: The processing flow of decomposable discrete cosine transform.

3.2 Decomposable Discrete Cosine Transform

Most of the existing frequency based deepfake detection methods extract frequency features from the whole image, which is non-sensitive to the local high-frequency information and heavily affects the detection of subtle forged faces. Inspired by ViT Dosovitskiy et al. (2020), the proposed Decomposable Discrete Cosine Transform method (DDCT) begins by splitting the image into small patches to extract the desired high-frequency features in a fine-grained manner.

The processing flow is shown in Fig.2, and we first transform the input RGB image $x^{rgb} \in R^{T\times C\times H}$ to the YCbCr space to eliminate the influence of image compression. Then we adopt the DCT convolution Qin et al. (2021) in the X^{YCbCr} space with the kernel size of 8×8 and the stride of 8, generating the local frequency feature $X^{freq} \in R^{T\times C\times \frac{H}{8}\times \frac{W}{8}}$, where the feature in the i^{th} row and j^{th} column reflects the delicate frequency clue of the small 8×8 patch in the original 224×224 image. To exploit more pinpoint frequency clues and eliminate redundant low-frequency information, we further separate high-frequency components x^{hf} from the whole frequency by employing a median filter with a smaller kernel size of 3 followed by a residual operation on each patch. Finally, we interpolate and resize the feature x^{hf} to the same size of X^{RGB}. In the patch-wise DDCT, we consider the frequency domain in a local region to explore more subtle artificial clues.

3.3 Short-term Embedding Module

The frame-level forgery methods inevitably trigger inconsistency between video frames. To explore the long-range video inconsistency while keeping the ability to perceive weak incoherency, the proposed ITSNet is consisted by two kinds of blocks: (1) the stacked SEM to extract the subtle local inconsistency among adjacent frames in the early stage and (2) the ultimate LEM to aggregate the global inconsistency.

The architecture of SEM is reflected in Fig. 3, where SEM incorporates two parallel branches to learn the subtle short-range inconsistency from a global and local perspectives. Compared with 3D

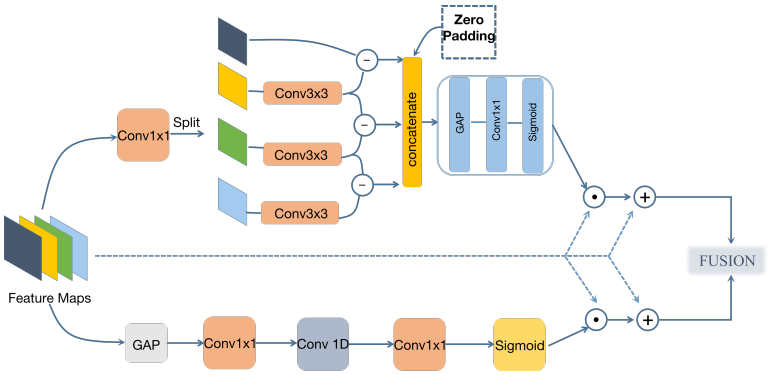

Figure 3: The architecture of the short-term embedding module.

convolution that can also encode the short-range dependency, the proposed SEM makes use of 2D and 1D convolution operations, which is not only effective but also computationally efficient through the local-global structure design.

The encoding of the global inconsistency is illustrated in the lower part of Figure 3. Given an input $X \in R^{T\times C\times h\times w}$, we first adopt the global average pooling (GAP) to aggregate the global spatial information. Then a symmetrical network is performed to squeeze and amplify the channel C with a ratio of r through stacked 1×1 convolution layers, which is similar to Channel Attention Module (CAM) Woo et al. (2018). However, different from the vanilla CAM, we creatively adapt a 1D convolution layer with a kernel size of 3 along the temporal dimension in the squeezed feature, which allows us to emphasize pertinent features in each frame according to the inter-channel relationship between adjacent frames. Finally, we apply a sigmoid function to acquire the channel attention map and generate the global enhanced inconsistency features X_g as described in Eq. 4.

$$X_g = X + \sigma\left(W_a\left(W_f\left(W_s\left(F^c_{avg}\right)\right)\right)\right) \cdot X, \tag{4}$$

where σ denoted the sigmoid function, $W_s \in R^{C/r\times C\times 1\times 1}$, $W_f \in R^{C/r\times C/r\times 3}$, and $W_a \in R^{C\times C/r\times 3}$.

To capture the fine-grained short-range temporal inconsistency, SEM targets to represent the local inconsistency among the adjacent frames at the same spatial position, as shown in the upper part of Figure 3. In particular, we first adopt a 1×1 convolution layer to squeeze the channel dimension by a ration of *r* (set to 16) and then obtain the local inconsistency $X_l \in R^{N\times 1\times \frac{C}{r}\times H\times W}$ through Eq. 5.

$$X_l(t) = M(X_r(t+1)) - X_r(t), \tag{5}$$

where M is a 3×3 convolution layer.

The representation X_l is concatenated along the time dimension and padded with zero to adapt to the original shape. To pay more attention to the temporal information, we employ a sequential GAP operation of 1×1 convolution layer to retain key channels by the scale ration *r*. Finally, the local attention representation is generated through the sigmoid function and the original features are calculated by Eq. 6.

$$X = \sigma(X_r(GAP(X_L))) + X \tag{6}$$

where X_L denotes the concatenation of X_l and $W_l \in R^{C\times H\times W}$ and X is the original input.

In summary, the subtle artifacts can be enhanced through the representation learning of the fine-grained local and global inconsistency among short-range frames. Since the inconsistency among adjacent frames mostly occurs at the texture level, we stack SEM modules in the early stage of the branch network and design the subsequent long-term embedding module to amplify and aggregate the overall temporal inconsistency from a long-range perspective.

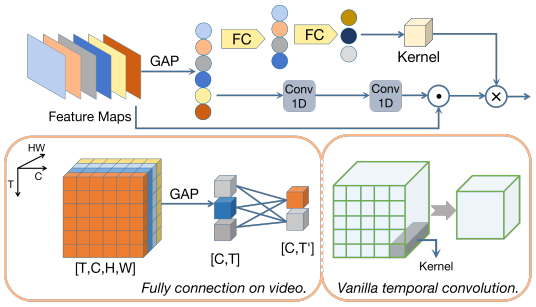

Figure 4: The architecture of the long-term embedding module.

3.4 Long-term Embedding Module

The complex temporal artificial clues contained in forgery videos can be further amplified in the long-range dependency. Motivated by this, we design LEM to assemble the temporal inconsistency from a long-range view.

Let $X \in R^{T\times C\times H\times W}$ denote the input frame set, LEM is proposed to aggregate the adjacent temporal artificial clues extracted by SEMs in a global temporal receptive field. As shown in Fig. 4, LEM consists of two parts, including (1) the enhancement branch on the global channel attention and (2) the adaptive temporal convolution mechanism. Specifically, we adopt a GAP to shrink the spatial dimension to $X_t \in R^{T\times C}$. The purpose is to eliminate redundant spatial information so as to concentrate on the temporal inconsistency.

In the enhancement branch with the global channel attention, we attempt to highlight channels related to the temporal inconsistency. To achieve the interaction between corresponding channels of video frames, we first transpose the temporal and channel dimension of X_t and then employ a 1D convolution stacked bottleneck layer with kernel size of 3 to generate the channel attention map. The enhanced feature $X_t^{'}$ can be acquired through the Hadamard product between the heatmap and the original feature.

To further emphasize the inconsistency between long-range frames, we design an adaptive temporal convolution mechanism to amplify and assemble the short-range temporal inconsistency based on the overall temporal inconsistency. We utilize two fully connected layers along the temporal dimension to extract the adaptive convolution kernel related to the inconsistency of the current video clips. Therefore, the LEM can fully explore the inconsistency between the global-range and local-range variation patterns, and further eliminate the negative influence of personal action habits and other task-irrelevant global information. Finally, the learned adaptive 3×3 convolution kernel is adopted on the enhanced representation $X_t^{'}$, which is formulated by Eq. 7 and Eq. 8.

$$K = fc_2(fc_1(F_{avg}(X_t))), \tag{7}$$

$$X = K(X_t^{'}), \tag{8}$$

where fc_1 and fc_2 are two fully connected layers, and K is the kernal of the adaptive temporal convolution.

4 Experiments

In this section, we first introduce the experiment setting including the implementation details, datasets, evaluation metrics, and baselines in Sec. 4.1. Then we conduct in-dataset and cross-dataset experiments in Sec. 4.2, where both qualitative and quantitative results are reported to demonstrate the effectiveness and generalization ability of the proposed ITSNet against numerous state-of-the-arts. Finally, ablation studies are executed in Sec. 4.3 to verify the effectiveness of components of the proposed ITSNet.

Methods	FF++(HQ)		FF++(LQ)		WildDeepfake	
	ACC	AUC	ACC	AUC	ACC	AUC
MesoNetAfchar et al. (2018)	83.10%	-	70.47%	-	-	64.47%
XceptionChollet (2017)	95.73%	-	86.86%	-	-	69.25%
Face X-rayLi et al. (2020b)	-	97 .80%	-	77 .30%	-	-
CNN-augWang et al. (2020)	96.90%	97 .20%	79.10%	78 .30%	-	-
F^3 -NetQian et al. (2020)	97.52%	98 .10%	90.43%	93.30%	80.66%	87.53%
Muti-AttZhao et al. (2021)	97.60%	99 .29%	88.69%	90 .40%	81.99%	90 .57%
PELGu et al. (2021)	97.63%	99 .32%	90.52%	94 .28%	84.14%	91 .62%
LipForensicsHaliassos et al. (2021)	98.80%	**99.70%**	94.20%	**98.10%**	-	-
FInferHu et al. (2022)	-	95 .67%	-	-	75.88%	81.83%
ITSNet (Ours)	**99.88%**	99.64%	**94.40%**	95.64%	**92.45%**	**92.34%**

Table 1: Quantitative comparisons on the FF++(HQ), FF++(LQ), and WildDeepfake datasets. The best performance is marked in bold.

4.1 Experimental Settings

Implementation details: We adopt the ResNet50 He et al. (2016) pre-trained on ImageNet as the backbone network of the proposed ITSNet. Based on the structure of ResNet, we first insert the SEM as the basic block in the first three stages of the backbone and replace the last stage with LEM. After each stage of the model, we adopt the interaction operation to achieve the cross-modality communication. With the sampling interval of 6 frames, we sample 8 frames from each video to construct the input of the model. For each frame, we first utilize S3FD Zhang et al. (2017) to detect and resize the face regions to 256×256, and then center crop it to 224×224. In the training stage, we utilize the AdamW optimizer and set the moment betas to (0.9, 0.999) and the learning rate to 5e-5. We train the network with a batch size of 4, and the total number of training epochs is set to 100 with a cosine learning rate schedule. In accordance with other methods, we also adopt the binary cross-entropy loss in the training phase. We implement the proposed method based on PytorchPaszke et al. (2017). The models are trained on two GeForce GTX 1080 GPUs.

Datasets: We evaluate our ITSNet on three widely used benchmarks, including FaceForensics++(FF++) Rossler et al. (2019), WildDeepfakeZi et al. (2020), and Celeb-DFLi et al. (2020c).

- **FaceForensic++** is the most commonly used deepfake benchmark that contains 1000 original videos collected from youtube, and 5000 correspondings are utilized to generate fake videos. FF++ employs five typical methods of face synthesis and manipulation, including Deepfakes (DF) git (a), Face2Face(F2F) Thies et al. (2016b), FaceSwap (FS) git (b), NeuralTextures (NT) Thies et al. (2019), and FaceShifter(Fsh) Li et al. (2019). Besides, FF++ provides compressed data in various degrees, including raw videos, high quality (HQ), and low quality (LQ).
- **WildDeepfake** is a real-world dataset that contains 7314 face sequences entirely, including 3805 natural and 3509 forged videos. These videos are widely spread on the Internet, with higher authenticity, more complex scenes, and more diverse facial gestures.
- **Celeb-DF** contains 590 real videos and 5639 composite videos. The real source videos are collected from publicly available youtube video clips with celebrities of diverse genders, ages, and ethnic groups.

Models for comparison: Our methods are compared with various state-of-the-art methods that can be divided into three categories. (1) *Vanllia CNN classifiers* including MesoNetAfchar et al. (2018), XceptionChollet (2017), and Resnet50 based CNN-aug Wang et al. (2020). (2) *Spatial-based methods* including Face X-rayLi et al. (2020b) which detects the blending boundary artifacts, F^3 -NetQian et al. (2020) that explores the artifacts in the frequency domain, PELGu et al. (2021) which eliminates irrelevant noise in the frequency domain, and Muti-AttZhao et al. (2021) which adopts a multi-spatial attention head to make the network focus on fine-grained local parts. (3) *Temporal-based methods* including LipforensicHaliassos et al. (2021) that exploits the temporal inconsistency of the lip reading, and FInfer Hu et al. (2022) that magnifies artifacts by predicting future frame.

Methods	Xception	Face X-ray	CNN-aug	F^3 -Net	Muti-Att	PEL	LipForensics	FInfer	**ITSNet (Ours)**
CD2	73.70%	79.50%	75.60%	69.75%	68.64%	69.18%	82.40%	70.60%	**85.97%**

Table 2: Cross-dataset evaluation on the CD2 dataset (AUC(%)).All methods are trained on FF++ (HQ) while tested on CD2.

Methods	Test					Avg
	DF	FS	F2F	NT	FSh	
XceptionChollet (2017)	93	51.2	86.8	79.7	72	76.6
CNN-augWang et al. (2020)	87.5	56.3	80.1	67.8	65.7	71.5
PatchForensicsChai et al. (2020)	94	60.5	87.3	84.8	65.7	78.5
CNN-GRUSabir et al. (2019)	97.6	47.6	85.8	86.6	80.8	79.7
Face X-rayLi et al. (2020b)	99.5	99.2	94.5	92.5	86.8	94.5
LipForensicsHaliassos et al. (2021)	**99.7**	90.1	**99.7**	99.1	97.1	97.1
ITSNet(Ours)	99.6	**99.6**	98.5	**99.6**	**99.2**	**99.3**

Table 3: Cross-dataset evaluation on the FF++ dataset (AUC(%)), which contains five manipulated methods (DF, FS, F2F, NT, Fsh). Data manipulated by four methods are used for training, while data tampered by the rest method are employed for testing.

4.2 Experimental Results

In-dataset Results: To demonstrate the effectiveness of the proposed ITSNet, we conduct in-dataset experiments on the most widely employed dataset FF++ and the most realistic dataset WildDeepfake. The experimental results are reported in Table 1. In general, ITSNet has achieved state-of-the-art results in both two datasets (FF++ and WildDeepfake) under various image qualities (HQ and LQ), which fully demonstrates the effectiveness of ITSNet.

In the FF++ dataset with high-quality (HQ), ITSNet achieves an accuracy of 99.88% and an AUC score of 99.64%, which outperforms the second best-performer Lipforensics Haliassos et al. (2021) by a gain of 1.08%. Lipforensics focus on the incoherence of mouth movements, while in our ITSNet the performance is further enhanced by the exploitation of long-range temporal inconsistency in two modalities.

In the setting of low-quality (LQ) in the FF++ dataset, ITSNet gains an ACC of 94.4%, which also occupies the first place among state-of-the-art methods. This also indicates that the cross-modality temporal inconsistency encoded by ITSNet is robust to video quality.

We also conduct experiments on the latest WildDeepfake dataset. As reflected in Table 1, our model reaches an AUC of 92.34% and ACC of 92.45%, which is decidedly superior to other state-of-the-art methods by a large margin. For instance, the second-best performer PEL Gu et al. (2021) obtains an ACC of 84.14%, which falls behind our ITSNet by a margin of 8.31%. Note that the video quality in WildDeepfake is similar to the real internet scenes where the most widespread forged videos have a complex posture and camera angles, which leads to a significant drop of many deepfake detectors. This also explains the effectiveness of our cross-modality inconsistency representation. Rather than focusing on spatial features, our ITSNet perceives the distribution of both domains over different time strides.

Cross-dataset Results: With the development of deep forgery technology, many mature and unseen forgery methods appear, which further brings challenges to the generalization ability of detection models. To demonstrate the generalization ability of the proposed ITSNet, we conduct cross-dataset experiments by training the model on FF++ (HQ) while evaluating it on CD2 via the metric of AUC. The experimental results are presented in Table 2. It can be seen that our ITSNet achieves the best generalization ability, with an AUC score of 85.97%, followed by LipForensics Haliassos et al. (2021) with an AUC of 82.4%.

To further examine the generalization ability of ITSNet on the data manipulated by unseen forgery methods, we follow the setting of Haliassos et al. (2021) to verify the generalized effect by the leave-one-out method. In particular, we train on arbitrary four types of forgery methods in FF++

Ablation Study	Methods	HQ	LQ
Components	RGB	96.78	91.07
	Freq	96.42	90.00
	RGB+Freq	97.28	92.64
	RGB+Freq+SEM	98.85	93.42
	RGB+Freq+SEM+LEM	99..28	93.85
Frequency domain	whole-frequency	98.50	94.78
	low-frequency	97.15	93.21
ITSNet (Ours)	RGB+Freq+SEM+LEM+Interaction (high-frequency)	**99.64**	**95.64**

Table 4: Abalation study of components and frequency domain on FF++(HQ) and FF++(LQ) dataset. Video-level AUC is reported.

and evaluate on the remaining one, which ensures that the model could not foresee the manipulated method in the test set during the training period. The results are shown in Table 3, among five evaluations on unseen forgery methods, our method attains three first places. Our approach obtains an average AUC of 99.3%, outperformed the second-best detector LipForensics Haliassos et al. (2021) by a large margin of 2.2%. This further demonstrates the excellent generalization performance of ITSNet, which is mainly benefit from the encoding of discriminate long-range temporal inconsistency.

4.3 Ablation Study

Effectiveness on the components. In this section, we validate the effectiveness of key components in our model. The ablation studies are conducted on both FF++(HQ) and FF++(LQ). To demonstrate the effectiveness and robustness of our ITSNet, we start with the baseline ResNet50 applied on the RGB stream and the frequency stream separately and progressively append the model with SEM, LEM, and interaction, as reported in Table 4.

It is clear that the proposed ITSNet achieves superior detection results. The adoption of two modalities obtains better performance than either one modality. The AUC results shown in the fourth row of Table 4 fully verify that subtle artificial clues can be exploited to determine deepfakes from the short-range temporal inconsistency. Likewise, the improvement obtained by adding LEM in the fifth row of Table 4 also proves that aggregating temporal correlations over long time strides facilitates the model to perceive unnatural motions. In the last row of Table 4, the proposed ITSNet boosts the detection performance by adding the interactions between two modalities, which further suggests the mutual promotion between the RGB representation and the fine-grained frequency representation.

Effectiveness on the frequency domain. To explore the effectiveness of the frequency representation, we design experiments for the DDCT proposed in Sec.3.2. Different frequency domains, including the entire-frequency domain, low frequency domain, and high frequency domain, are represented in the frequency stream of ITSNet and are examined by the accuracy of the deepfake detection task. The low-frequency component is generated by employing a median filter operation on the entire-frequency representation learned from the DCT convolution while taking the residual as the high-frequency component as described in Sec. 3.2. The frequency ablation experiments are conducted on both FF++(HQ) and FF++(LQ) with the same parameter settings, and the results are illustrated in Table 4. It can be concluded that the distribution of high-frequency components is more effective for detecting face forgery videos.

5 Conclusion

In this paper, we propose a novel end-to-end forgery video detection method, ISTNet. ITSNet is constructed by an interactive two-stream network to encode both RGB and frequency artificial clues for the real/fake binary classification. To extract discriminate high-frequency components, we design a patch-wise decomposable DCT transform (DDCT) in a fine-grained way. To exploit inconsistency representation encoding both short-term and long-range dependencies, the short-term embedding module (SEM) and long-term embedding module (LEM) are developed. Extensive experiments are conducted and experimental results demonstrate that our model outperforms the other state-of-the-art methods in both in-dataset and cross-dataset evaluations.

REFERENCES

Deepfakes. https://github.com/deepfakes/faceswap., a. Accessed: 2018-10-29.

Faceswap. https://github.com/MarekKowalski/FaceSwap/, b. Accessed: 2018-10-29.

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7. IEEE, 2018.

Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 660–661, 2020.

J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. *IEEE*, 2018.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, pp. 103–120. Springer, 2020.

Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection, 2021. URL https://arxiv.org/abs/2105.02577.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pp. 3247–3258. PMLR, 2020.

Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. *arXiv preprint arXiv:2112.13977*, 2021.

David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pp. 1–6. IEEE, 2018.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 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, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang. Learning to detect fake face images in the wild. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pp. 388–391. IEEE, 2018.

Juan Hu, Xin Liao, Jinwen Liang, Wenbo Zhou, and Zheng Qin. Finfer: Frame inference-based deepfake detection for high-visual-quality videos. *IEEE Trans. Pattern Anal. Mach. Intell*, pp. 1–9, 2022.

I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. *IEEE*, 2017.

H Li, B Li, S Tan, and J Huang. Detection of deep network generated images using disparities in color components. arxiv 2018. *arXiv preprint arXiv:1808.07276*.

L. Li, J Bao, H Yang, D. Chen, and F. Wen. Advancing high fidelity identity swapping for forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

L. Li, J. Bao, T. Zhang, H. Yang, and B. Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

Y. Li, M. C. Chang, and S. Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 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, 2020c.

Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: Face swapping and editing using face and hair representation in latent spaces. *ACM*, 2018.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pp. 86–103. Springer, 2020.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 783–792, 2021.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019.

Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3 (1):80–87, 2019.

J. Thies, M Zollhöfer, M. Stamminger, C. Theobalt, and M Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2016a.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387–2395, 2016b.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In-So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

Li Yuezun Lyu Siwei Yang, Xin. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265. IEEE, 2019.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pp. 192–201, 2017.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2019.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2185–2194, 2021.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15044–15054, 2021.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2382–2390, 2020.