Delving into the Local: Dynamic Inconsistency Learning for DeepFake Video Detection

Zhihao Gu[1,2*], **Yang Chen**[2*], **Taiping Yao**[2*], **Shouhong Ding**[2†], **Jilin Li**[2], **Lizhuang Ma**[1,3,4†]

[1]School of Electronic and Electrical Engineering, Shanghai Jiao Tong University,
[2]YouTu Lab, Tencent
[3]MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[4]East China Normal University
ellery-holmes@sjtu.edu.cn, {wizyangchen, taipingyao, ericshding, jerolinli}@tencent.com, ma-lz@cs.sjtu.edu.cn

Abstract

The rapid development of facial manipulation techniques has aroused public concerns in recent years. Existing deepfake video detection approaches attempt to capture the discriminative features between real and fake faces based on temporal modelling. However, these works impose supervisions on sparsely sampled video frames but overlook the local motions among adjacent frames, which instead encode rich inconsistency information that can serve as an efficient indicator for DeepFake video detection. To mitigate this issue, we delves into the local motion and propose a novel sampling unit named *snippet* which contains a few successive videos frames for local temporal inconsistency learning. Moreover, we elaborately design an Intra-Snippet Inconsistency Module (Intra-SIM) and an Inter-Snippet Interaction Module (Inter-SIM) to establish a dynamic inconsistency modelling framework. Specifically, the Intra-SIM applies bi-directional temporal difference operations and a learnable convolution kernel to mine the short-term motions within each snippet. The Inter-SIM is then devised to promote the cross-snippet information interaction to form global representations. The Intra-SIM and Inter-SIM work in an alternate manner and can be plugged into existing 2D CNNs. Our method outperforms the state of the art competitors on four popular benchmark dataset, i.e., FaceForensics++, Celeb-DF, DFDC and WildDeepfake. Besides, extensive experiments and visualizations are also presented to further illustrate its effectiveness.

Introduction

With the rapid development of deep learning-based methods, especially generative models, various DeepFake techniques (Koujan et al. 2020; Nirkin, Keller, and Hassner 2019) have been proposed. Since these techniques can synthesize more and more realistic DeepFakes that are hardly distinguishable by humans, abuse of them can easily trigger severe societal problems or political threats over the world. Therefore, it is of great importance to develop effective methods for face forgery detection.

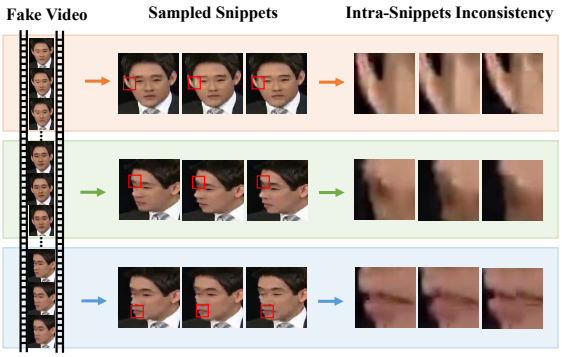

Figure 1: Illustration of local inconsistency in snippets. Since the snippet contains densely sampled frames, different type of inconsistency caused by subtle motions can be captured better, which can serve as a strong indicator for identifying DeepFakes.

Recently, significant progress has been achieved in DeepFake detection. As the release of large-scale face forgery video datasets (Rossler et al. 2019; Li et al. 2020b; Dolhansky et al. 2019; Zi et al. 2020), it enables the training of deep convolutional neural networks (DCNNs) to identify DeepFakes with various manipulations, *e.g.,* DeepFakes and FaceSwap. Image-based methods focus on mining various frame-level cues including frequency information (Qian et al. 2020), auxiliary masks (Chen et al. 2021; Wang et al. 2020) and textural information (Zhao et al. 2021) to improve performance. However, when encountering extremely realistic images, image-based methods may fail to mine them and thus have limited performance. Besides, they do not consider the inconsistent facial movements between real and fake videos, which derives from the frame-by-frame manipulation.

Therefore, many researchers recently develop video-based methods to capture such inconsistency as a discriminative clue for DeepFake detection. Earlier methods treat

*These authors contributed equally. This work was done when Zhihao Gu was a research intern at Tencent YouTu Lab.
†Corresponding authors.
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

this task as a general temporal modelling problem and classic approaches like LSTM (Sohrawardi et al. 2019) and 3DCNN (Zi et al. 2020) are applied to solve this problem. However, they are not specifically designed for DeepFake detection and accordingly achieve inferior performance, not to mention their high computational cost. More recent works start to study the inconsistency to locate the forgery trace in DeepFake videos, *e.g.,* S-MIL (Li et al. 2020a) and STIL (Gu et al. 2021), and show promising results. The state-of-the-art STIL observes that the motion between adjacent frames in real videos is more smooth than fake ones. They term this clue as a form of inconsistency and exploit the temporal difference over adjacent frames to model it. However, they apply a sparse sampling strategy for each video and the interval of sampled frames might be too big to capture this inconsistency resulting from subtle motion.

In our perspective, the inconsistency caused by frame-by-frame manipulation can be observed best within densely sampled frames, *i.e.,* the temporal locality plays a key role for inconsistency mining, as depicted in Fig 1. Therefore, we propose to extract inconsistency information based on video snippets, each of which is composed of N successive video frames. All snippets span uniformly over the entire video to form a local-to-global view. To deal with these snippets, an Intra-Snippet Inconsistency Module (Intra-SIM) and an Inter-Snippet Interaction Module (Inter-SIM) are devised and they are performed in an alternate manner. Specifically, the Intra-SIM first adopts bi-directional temporal difference operation to model the intra-snippet inconsistency. Then, a novel coordinate attention over $H \times T$ and $T \times W$ dimensions are respectively exploited to extract fine-grained while more comprehensive representations. Finally, learnable kernels are introduced to adaptively aggregate the intra-snippet inconsistency information. The Inter-SIM employs a new two-branch structure to establish a cross-snippet view for interaction promotion. Both of the Intra-SIM and Inter-SIM serve as plug-and-play modules and could be integrated into the off-the-shelf 2D CNNs.

The proposed method surpasses the state-of-the-art competitors on both intra-dataset evaluation, *i.e.,* FF++ (Rossler et al. 2019), DFDC (Dolhansky et al. 2019), Celeb-DF (Li et al. 2020b) and WildDeepfake (Zi et al. 2020) datasets, and inter-dataset generalization settings. The visualization experiment further demonstrates the effectiveness of each component. Interestingly, when encountering partially forged videos, the inter-snippet motion activation map in Inter-SIM can correctly localize the forged faces in both short and long-term video sequences. In summary, our main contributions are three-folds:

- We propose a novel DeepFake video detection scheme by focusing on mining local inconsistency encoded in video snippets, which contain a few successive video frames.
- A novel Intra-SIM is devised to learn snippet-specific short-term inconsistency and a new Inter-SIM is designed to help snippets better interact with each other. Both of them work as plug-and-play modules and can be easily integrated with the existing 2D backbones.
- We set a new state-of-the-art result on four popular benchmarks, *i.e.,* FaceForensics++, Celeb-DF, DFDC and WildDeepfake datasets. Cross-dataset generalization and visualizations further validate the effectiveness of the proposed method.

Related Work

DeepFake Detection. The existing deep learning-based forgery detectors can be classified into image and video-based methods. The image-based methods aim to mine discriminative frame-level representations for identification. (Rossler et al. 2019) evaluates five well-known network architectures to solve the task. (Dang et al. 2020) proposes a weakly supervised methods to highlight the informative regions for processing and improving the feature maps for classification. (Li et al. 2021) introduces a frequency-aware feature learning framework, which compresses intra-class variations of real faces and enlarges inter-class differences. All these methods achieve impressive performance in image-level detection. However, as the develop of manipulation techniques, the frame-level forgery trace can hardly be captured. Recent works treat this task as a video-level representation learning problem and most of efforts are devoted to modeling the inconsistency presented in real and fake videos for classification. (Sabir et al. 2019) achieves state-of-the-art performance through combining recurrent convolutional strategies along with face pre-processing techniques. (Qi et al. 2020) reveals forgery by predicting the heartbeat rhythms from videos since the rhythms will be broken by manipulations. (Masi et al. 2020) presents a two-branch architecture to amplify artifacts and suppress high-level facial contents for isolating Deepfakes. (Li et al. 2020a) introduces the multiple instance learning framework for the partial face attack in videos. (Haliassos et al. 2021) targets the semantic irregularities in mouth movement presented in fake videos for better cross-dataset generalization. (Agarwal and Farid 2021) describes a forensic technique that exploits the abnormal shape of the ear and ear canal caused by movement of lower jaw. Actually, these methods model the long-term inconsistency and the short-term inconsistency is completely ignored, which, in this paper, we consider extremely crucial to the task.

Video Analysis. The core of video-related tasks is temporal modeling and many efforts are devoted to modeling the temporal dependency. Early works such C3D (Tran et al. 2014) and I3D (Carreira and Zisserman 2017) exploits the 3D CNNs for temporal modeling. These models either are computationally expensive than the 2D counterparts or learn spatial-temporal features from snippets without considering the video-level evolution. To mitigate the efficiency issue, several efficient modules are proposed to equip the 2D CNNs with the capacity of temporal modeling. (Lin, Gan, and Han 2019) proposes temporal shift module to shift part of channels along temporal dimension for efficient temporal modeling. (Wang et al. 2021) establishes an two-level temporal modeling paradigm to capture both short and long-term information over the entire video. To capture the temporal evolution, (Zhang et al. 2020) presents a novel video-level 4D convolution (V4D) with residual connections to simultaneously capture the long-term relation between

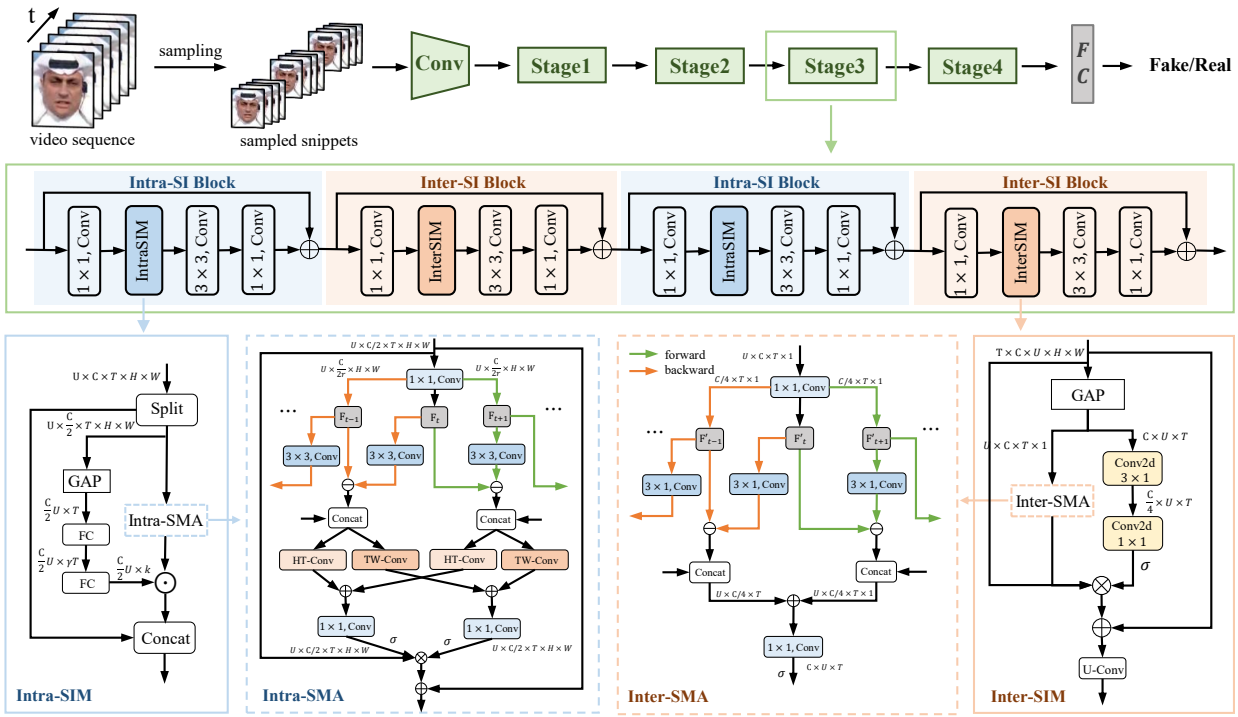

Figure 2: The overall architecture of the proposed method. We sample several snippets uniformly from each video and each snippet contains a few successive frames. In ResNet backbones, the proposed Intra-SIM and Inter-SIM modules are inserted to res-blocks in an alternate manner and turn them into the Intra-SI and Inter-SI blocks. HT-Conv and TW-Conv represent convolutions over $T \times W$ and $T \times W$ dimensions separately. $\oplus$, $\ominus$, $\otimes$ and $\odot$ denote element-wise addition, subtraction, multiplication and depth-wise convolution, respectively.

snippets while maintaining 3D features before interaction. (Wu et al. 2021) generates dynamic convolutional kernels to adaptively aggregate long-range temporal information from adjacent snippets. These methods focus on modeling long-range dependency and applying them without task-specific knowledge of fake videos may limit their performance.

Different from all the methods mentioned above, we focus on short-term motion modeling and argue that, for DeepFake video detection, the key lies in modeling the intra-snippet inconsistency. Besides, promoting the interaction between snippets can help learn better video-level representations.

Approach

In this section, we first present the technical details of the proposed Intra-Snippet Inconsistency Module (Intra-SIM) and the Inter-Snippet Interaction Module (Inter-SIM). Then we describe how to instantiate them with the off-the-shelf 2D CNNs.

Intra-Snippet Inconsistency Module

As already mentioned, learning from short-term motion plays a vital role in mining temporal inconsistency. To this end, we sample the video sequence uniformly into U snippets, each of which contains T successive frames rather than a single frame like the previous works do. Our designed Intra-Snippet Inconsistency Module (Intra-SIM) then takes frames within each snippet to model the local inconsistency encoded in subtle motions. As illustrated in Fig. 2, our Intra-SIM works in a two-branch manner. The short branch is a residual connection so that the original representations is directly accessible. The long branch consists of an Intra-Snippet Motion Attention (Intra-SMA) module and a pathway with learnable convolutional kernels to adaptively aggregate intra-snippet inconsistency information.

In consideration of computation efficiency, let tensor $I \in R^{C \times T \times H \times W}$ denotes the input feature within each snippet, where C is the number of channels and T, H, W are the temporal and spatial dimensions. We first split I into two equal parts along the channel dimension to get I_1 and I_2, and then feed them into subsequent branches. To model temporal relation, Intra-SMA applies bi-directional temporal difference guided coordinate attention to make the network attend to local motions. Take the forward flow in Fig. 2 for example, the input $I_2 = [F_1, \cdots, F_T] \in R^{\frac{C}{2} \times T \times H \times W}$ is first compressed by a ratio r and then used to calculate the temporal difference among adjacent frames:

$$D_{t,t+1} = F_t - \text{Conv}_{3\times3}(F_{t+1}), \tag{1}$$

where $D_{t,t+1}$ represents the forward temporal difference for F_t and $\text{Conv}_{3\times3}$ is channel-wise convolution. After that, $D_{t,t+1}$ is reshaped into two coordinate-wise representations,

Methods	FaceForensics++ HQ				FaceForensics++ LQ			
	DF	F2F	FS	NT	DF	F2F	FS	NT
ResNet-50	0.9893	0.9857	0.9964	0.9500	0.9536	0.8893	0.9464	0.8750
Xception	0.9893	0.9893	0.9964	0.9500	0.9678	0.9107	0.9464	0.8714
LSTM	0.9964	0.9929	0.9821	0.9393	0.9643	0.8821	0.9429	0.8821
C3D	0.9286	0.8857	0.9179	0.8964	0.8929	0.8286	0.8786	0.8714
I3D	0.9286	0.9286	0.9643	0.9036	0.9107	0.8643	0.9143	0.7857
TEI†	0.9786	0.9714	0.9750	0.9429	0.9500	0.9107	0.9464	0.9036
DSANet†	0.9929	0.9929	0.9964	0.9571	0.9679	0.9321	0.9536	0.9178
V4D†	0.9964	0.9929	0.9964	0.9607	0.9786	0.9357	0.9536	0.9250
FaceNetLSTM	0.8900	0.8700	0.9000	-	-	-	-	-
Co-motion-70	0.9910	0.9325	0.9830	0.9045	-	-	-	-
DeepRhythm	0.9870	0.9890	0.9780	-	-	-	-	-
ADDNet-3d†	0.9214	0.8393	0.9250	0.7821	0.9036	0.7821	0.8000	0.6929
S-MIL	0.9857	0.9929	0.9929	0.9571	0.9679	0.9143	0.9464	0.8857
S-MIL-T	0.9964	**0.9964**	1.0	0.9429	0.9714	0.9107	0.9607	0.8679
STIL	0.9964	0.9929	1.0	0.9536	0.9821	0.9214	0.9714	0.9178
Ours	**1.0**	0.9929	**1.0**	**0.9643**	**0.9928**	**0.9571**	**0.9786**	**0.9428**

Table 1: Comparison with the state-of-the-art DeepFake detectors on FF++ dataset. 'LQ' means low image quality while 'HQ' stands for high quality. † implies re-implementation. Best results are bold in font and '-' indicates results are unavailable.

i.e., $D^h_{t,t+1} \in R^{W\times\frac{C}{2}\times H\times T}$ and $D^w_{t,t+1} \in R^{H\times\frac{C}{2}\times T\times W}$, which further undergo through a multi-scale structure to capture fine-grained short-term motion information:

$$D^H_{t,t+1} = \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(D^h_{t,t+1}) + D^h_{t,t+1}), \quad (2)$$

$$D^W_{t,t+1} = \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(D^w_{t,t+1}) + D^w_{t,t+1}), \quad (3)$$

where $D^H_{t,t+1}$, $D^W_{t,t+1}$ and $\mathrm{Conv}_{1\times1}$ are forward vertical inconsistency, forward horizontal inconsistency and 1×1 convolution for dimension recovery, respectively. Backward vertical difference $D^H_{t+1,t}$ and backward horizontal difference $D^W_{t+1,t}$ can be obtained in a similar way. Averaging these features and applying a sigmoid function, the horizontal and vertical attention $Atten_H$ and $Atten_W$ can be obtained. As shown in Fig. 2, to adaptively aggregate the intra-snippet inconsistency information, we design to automatically learn a 1D convolution kernel which is appied to $Atten_H$ and $Atten_W$. In this learning process, we first exploit a global average pooling (GAP) operation to squeeze the spatial dimension for global view, and then two fully connected layers, *i.e.,* $\phi_1 : R^T \to R^{\gamma T}$ and $\phi_2 : R^{\gamma T} \to R^k$ are performed, finally a softmax operation comes up. This process can be modelled by the following formulation:

$$\mathcal{K}(X_2) = \mathrm{softmax}(\phi_2 \circ \delta \circ \phi_1(\mathrm{GAP}(X_2))) \quad (4)$$

where $\circ$ is for function composition and δ denotes ReLU activation function.

Once obtaining the intra-SMA and the kerneks, the intra-snippet inconsistency can be formulated:

$$O_2 = \mathcal{K}(X_2) \otimes (Atten_h \odot Atten_w \odot X_2 + X_2), \quad (5)$$

where $\otimes$ represents depth-wise convolution and $\odot$ stands for element-wise multiplication. Finally, we get the output:

$$O_{Intra} = \mathrm{Concat}[I_1, O_2]. \quad (6)$$

where Concat represents the concatenation operation along the channel dimension.

Inter-Snippet Interaction Module

The Intra-SIM adaptively captures intra-snippet inconsistency. However, such representation only contains the temporally local information and the relation between snippets is ignored, which is also important. Therefore, our Inter-Snippet Interaction Module (Inter-SIM) focuses on promoting the interaction across snippets from a global view to enhance the representation via a novel structure with different kind of interaction modeling, as shown in Fig. 2.

Formally, let tensor $F \in R^{T\times C\times U\times H\times W}$ be the module input. It is first processed by GAP to obtain a global representation $\bar{F} \in R^{C\times U\times T}$ and then passed through a two-branch structure for different interaction modeling. These two branches are complementary to each other in terms of intra-snippet information. Among them, one branch directly captures the inter-snippet interaction without introducing intra-snippet information:

$$\bar{F}_1 = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{BN}(\mathrm{Conv}_{3\times1}(\bar{\mathrm{F}})))) \quad (7)$$

where $\mathrm{Conv}_{3\times1}$ is spatial convolution with kernel size 3×1 for snippet-wise feature extraction and dimension reduction, and $\mathrm{Conv}_{1\times1}$ stands for convolution with size 1×1 for dimension recovery. The other branch inter-snippet motion attention, which is designed to be computationally efficient while containing a larger intra-snippet fields-of-view. Given the feature $\hat{F} \in R^{\frac{C}{r}\times U\times T}$ processed by the squeeze operation $\mathrm{Conv}_{1\times1}$ from $\bar{F}$, the intra-snippet interaction is first captured by $\mathrm{Conv}_{1\times3}$ and then the bi-directional facial movements are modeled in a similar way to Eq. (1):

$$\hat{D}_{u,u+1} = \hat{F}_u - \mathrm{Conv}_{1\times3}(\hat{F}_{u+1}), \quad (8)$$

$$\hat{D}_{u+1,u} = \hat{F}_{u+1} - \mathrm{Conv}_{1\times3}(\hat{F}_u). \quad (9)$$

Therefore, we define the inter-snippet information with intra-snippet interaction as:

$$\bar{F}_2 = \sigma(\mathrm{Conv}_{1\times1}(\hat{D}_{u,u+1} + \hat{D}_{u+1,u}), \quad (10)$$

Methods	Celeb-DF	DFDC	WildDeepFake
Xception	0.9944	0.8458	0.8325
I3D†	0.9923	0.8082	0.6269
D-FWA	0.9858	0.8511	-
DIANet	-	0.8583	-
TEI†	0.9912	0.8697	0.8164
V4D	0.9942	0.8739	0.8375
DSANet†	0.9942	0.8867	0.8474
ADDNet-3D†	0.9516	0.7966	0.6550
S-MIL	0.9923	0.8378	-
S-IML-T	0.9884	0.8511	-
STIL†	0.9961	0.8980	0.8462
Ours	**0.9961**	**0.9279**	**0.8511**

Table 2: Comparison on Celeb-DF, DFDC, and WildDeepfake datasets. † implies our implementation.

Methods	FF++ DF	Celeb-DF	DFDC
Xception	0.9550	0.6550	0.5939
I3D†	0.9541	0.7411	0.6687
VA-LogReg	0.7800	0.5510	-
TEI†	0.9654	0.7466	0.6742
D-FWA	0.8100	0.5690	-
Capsule	0.9660	0.5750	-
V4D†	0.9674	0.7008	0.6734
DIANet	0.9040	0.7040	-
DSANet†	0.9688	0.7371	0.6808
DoubleRNN	0.9318	0.7341	-
ADDNet-3D†	0.9622	0.6085	0.6589
STIL†	0.9712	0.7558	0.6788
Ours	**0.9819**	**0.7765**	**0.6843**

Table 3: Comparison on cross-dataset generalization in terms of AUC. † implies re-implementation.

Finally, the representation power of temporal convolution $\mathrm{Conv_U}$, with size 3×1, is enhanced as follows:

$$O_{inter} = \mathrm{Conv_U}(\bar{F}_1 \odot \bar{F}_2 \odot F + F). \tag{11}$$

where each position of F is aware of various information.

Instantiation

The proposed method is instantiated with the well-known ResNet-50 (He et al. 2016) in light of its trade-off between accuracy and speed. We insert Intra-SIM and Inter-SIM right before the spatial convolution in each resnet block to form the Intra-snippet Inconsistency Block (Intra-SI Block) and Inter-Snippet Interaction Block (Inter-SI Block), as demonstrated in Figure 2. Unless specified, they are placed in an alternate manner.

Experiments

Experimental Settings

Datasets. We evaluate our method on four widely used benchmarks: FaceForensics++ (Rossler et al. 2019), DFDC (Dolhansky et al. 2019), Celeb-DF (Li et al. 2020b) and WildDeepfake (Zi et al. 2020).

- **FaceForensics++** contains multiple video quality, *e.g.,* high quality (HQ) with nearly no visual loss and low quality (LQ), which is visually blurry. Each of them consists of 1,000 real and 4,000 fake videos generated from four forgery techniques, *i.e.,* DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT).
- **DFDC** makes up around 5,000 videos with several unknown manipulation methods. Since faces in the video may be partially forged, state-of-the-art detectors perform not very well on this challenging dataset.
- **Celeb-DF** is comprised of 590 real videos and 5,639 forged videos totally from publicly available YouTube video clips. An improved synthesis process is used to improve various visual artifacts presented in these videos.
- **WildDeepfake** is a real-world face forgery dataset and consists of 7,314 face sequences purely collected from the Internet. These videos are crafted by forgery methods of different type and thus present diverse. The duration of each video varies a lot and thus is more challenging.

Baseline Methods. To demonstrate the effectiveness of the proposed method, we compare it with several representative works in face forgery detection and video analysis. For image-based methods, ResNet (He et al. 2016), Xception (Rossler et al. 2019) and VA-LogReg (Matern, Riess, and Stamminger 2019) are chosen and video-level results are averaged from frame-level predictions. For video-based DeepFake detectors, state-of-the-art D-FWA (Li and Lyu 2019), FaceNetLSTM (Sohrawardi et al. 2019), Capsule (Nguyen, Yamagishi, and Echizen 2019), Co-motion (Wang, Zhou, and Wu 2020), S-MIL (Li et al. 2020a), DeepRhythm (Qi et al. 2020), ADDNet-3d (Zi et al. 2020), STIL (Gu et al. 2021) and DIANet (Hu et al. 2021) are selected. What's more, action recognition models including LSTM (Hochreiter and Schmidhuber 1997), C3D (Tran et al. 2014), I3D (Carreira and Zisserman 2017), TEI (Liu et al. 2020), V4D (Zhang et al. 2020) and DSANet (Wu et al. 2021) are adopted to illustrate the superiority of ours.

Implementation Details Following the common practice (Li et al. 2020a), we use dlib to detect face for FF++ dataset as data pre-processing, while MTCNN (Zhang et al. 2016) is exploited for other datasets. The ImageNet (Deng et al. 2009) pre-trained 2D ResNet-50 (He et al. 2016) is used as our backbone and both Intra-SIM and Inter-SIM are randomly initialized. All snippets are sampled uniformly from each video sequence and we sample $U = 4$ snippets each with $T = 4$ frames. The image is resized to 224×224 during training. We adopt the Adam (Kingma and Ba 2014) as optimizer to optimize the binary cross-entropy loss. The batch size is 10 and the initial learning rate is 10^{-4}. The total epoch is 30 for all datasets and 45 for cross-dataset generalization. We divide the learning rate by 10 when the performance on validation set saturates. Only horizontal flip is employed for augmentation. During inference, we sample

Intra	Inter	DF	F2F	FS	NT
		0.9536	0.8893	0.9464	0.8750
√		0.9750	0.9250	0.9643	0.9214
	√	0.9714	0.9214	0.9607	0.9036
√	√	**0.9928**	**0.9571**	**0.9786**	**0.9428**

Table 4: Study on effects of Intra-SIM and Inter-SIM. We insert the corresponding modules into all stages.

BP	CA	LK	DF	F2F	FS	NT
			0.9750	0.9107	0.9571	0.9143
√			0.9821	0.9286	0.9643	0.9214
√	√		0.9928	0.9393	0.9678	0.9286
√	√	√	**0.9928**	**0.9571**	**0.9786**	**0.9428**

Table 5: Study on impacts of bi-direction path (BP), coordinate attention (CA) and learnable kernels (LK).

Stages	DF	F2F	FS	NT
Sateg_{1-2}	0.9857	0.9500	0.9607	0.9357
Sateg_{2-3}	0.9821	0.9286	0.9607	0.9214
Sateg_{3-4}	0.9786	0.9286	0.9607	0.9071
Sateg_{2-4}	0.9893	0.9393	0.9464	0.9107
Sateg_{1-3}	0.9893	0.9464	0.9678	0.9357
Ours	**0.9928**	**0.9571**	**0.9786**	**0.9428**

Table 6: Study on locations of Intra-SIM and Inter-SIM. We place them at different stages of ResNet-50 and they are put in a alternate manner in each stage.

$U = 8$ snippets with $T = 4$ frames and resize them into the same size as in training.

Intra-dataset Comparisons

In this section, we perform comparisons on four widely used benchmarks, *i.e.,* FF++, Celeb-DF, DFDC and WildDeepfake, to evaluation model effectiveness in terms of accuracy.

Results on FF++. We conduct comprehensive experiments on FF++ dataset under both low quality (LQ) and high quality (HQ) image qualities and report comparisons against state-of-the-art works in Table 1. It is clear that: (1) Advanced video-based action recognition models have better results than image-based methods. This is reasonable as the temporal motion information is helpful in DeepFake video detection. Besides, the performance of V4D is still competitive with the state-of-the-art STIL which is video-based and specifically designed to extract temporal inconsistency in DeepFake videos. The reason behind is that methods inlcuding ADDNet-3d, S-MIL and STIL all employ the sparse sampling strategy and do not delve into the intra-snippet information, which leads to limited capacity of mining inconsistency. (2) Our method utilizes the Intra-SIM to grasp the local inconsistencies caused by subtle motion, and further promotes the cross-snippet interaction by Inter-SIM to form a global view. Therefore, our method outperforms nearly all compared opponents on all settings except for F2F HQ, which is slightly worse than S-MIL-T. Moreover, on the most challenging NT LQ setting, we achieve 94.28% accuracy, exceeding 1.78% than the best action recognition model V4D and 2.5% than the state-of-the-art DeepFake detection method STIL. In addition, when transferred to C40 setting with lower image quality, all methods encounter severe performance drop except for our method. This demonstrates the necessity of modelling local inconsistency and the effectiveness of our method.

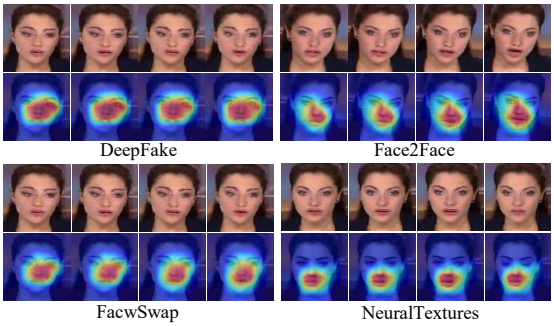

Figure 3: CAM of model outputs against four manipulations in FF++ dataset. For simplicity, we only visualize the CAM on frames within snippets.

Results on Celeb-DF, DFDC and WildDeepfake. We also evaluate our method on other three popular datasets, *i.e.,* Celeb-DF, DFDC and WildDeepfake datasets, as listed in Table 2. It shows that our method outperforms all the competitors, especially on DFDC by a large margin of 4.19% than STIL. This is because DFDC dataset contains many partially forged videos, where sparsely sampled frames might not be able to cover the forged motions and the inconsistency within. However, our method is built on snippets and the learnable kernels enrich each feature with dynamically short-term motion information. The Inter-SIM further promotes the interaction between snippets for long-term representations.

Inter-dataset Comparisons

Results on cross-dataset generalization. Following (Masi et al. 2020), we train the model on FF++ LQ datasets against four manipulations and perform cross-dataset tests on FF++ DF, Celeb-DF and DFDC datasets in verification of model generalization. Comparisons under AUC metrics are shown in Table 3. It shows that our method still suppresses all the compared competitors and achieves 2.07% and 0.35% performance gains than state-of-the-art results. It's also noticeable that the frame-based Xception has severe performance drop on unseen datasets. Since the frame-based methods mainly focus on image-level forgery patterns but neglect the temporal inconsistencies, they are more prone to overfitting. Compared to all the image- and video-based counterparts, our method delves into the inconsistency encoded in local motions and develops novel intra- and inter-inconsistencies mining scheme, which can better grasp the forgery nature in DeepFake videos. Therefore, our method has superior generalization ability.

Ablation Study

In this section, we conduct comprehensive ablation studies on FF++ LQ part to explore the effectiveness of each com-

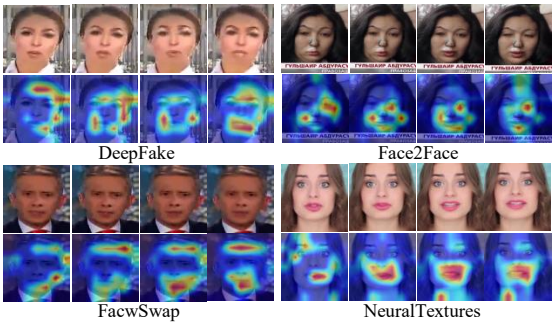

Figure 4: Snippet CAM of Intra-SMA against four manipulations. Samples are from FF++ dataset.

ponent in the proposed modules from Tab. 4 to Tab. 6.

Study on effects of Inter-SIM and Intra-SIM. As shown in Table 4, we study the impacts of Intra-SIM and Inter-SIM. Without introducing temporal information, the vanilla ResNet-50 has poor performance. Inserting only Inter-SIM or Intra-SIM already improves the performance a lot (at least 1.5% on each category). Obviously, both inter and inter-snippet information are vital and the combination of them boosts the performance. Note that the intra-snippet information contributes to the improvement more, which again illustrates the importance of intra-snippet dynamics.

Study on elements in Inter-SIM. The Inter-SIM contains three key parts, *i.e.,* bi-directional path (BP), coordinate attention (CA) and learnable convolution kernels (LK) and the results are listed in Table 5. The first row in the table demonstrates that without the guidance from fine-grained motion information, its overall performance is worse especially on NT (91.43%). Based on it, three key elements are gradually added. Among them, using learnable kernels gives the largest increment than other elements (about 4% on NT, 1% on FS and 2% on F2F). There is no doubt that best results benefits from exploiting all of them.

Study on different locations. We study the locations where to insert the Intra-SIM and Inter-SIM modules and the corresponding result are listed in Table 6. As can be observed that early stage (*i.e.,* stage 1-2) performs consistently better than middle (*i.e.,* stage 2-3) and late stages (*i.e.,* stage 3-4). More importantly, intra and inter-snippet information in low-level representations, *i.e.,* stage 1, seem to be more important than others (see the 4th and 5th rows). Of course, inserting them into all stages performs best.

Visualization Analysis

We adopt the Grad-CAM (Selvaraju et al. 2017) to visualize the class activation maps, Intra-SMA and U-T attention maps in Inter-SIM from Figure 3-5.

Class activation maps. The activation maps against four manipulation techniques in FF++ dataset are visualized in Figure 3. Both learnt maps for DeepFakes and FaceSwap focus on a large center regions whereas on Face2Face and NeuralTextures they shrink to the forged areas. For example, the NeuralTextures mainly focuses on forging the mouth areas. This is interesting as the proposed method is still able

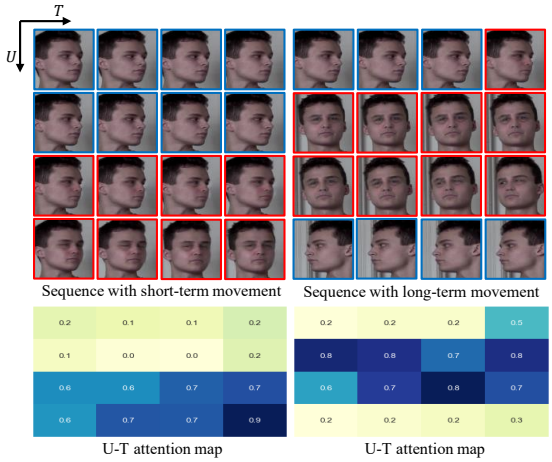

Figure 5: Visualization of U-T attention map in Inter-SIM on both short and long-term movement sequences. Red boxes correspond to fake faces and numbers in U-T map represent the fake probability.

to learn the essential features under only video-level supervisions.

Short-term motion activation maps. In order to demonstrate what the model learns from the short-term motions in each snippet, we visualize the activation maps of intra-SMA in Figure 4. Since DeepFakes and FaceSwap manipulate with the whole facial area, the activated maps indeed land on the face contours. Activations for Face2Face focus on facial expressions while activations for NeuralTextures are responsible to mouth movements. Moreover, the attended locations vary along with time, possibly seeking the most salient forgery traces in the short-term motions.

U-T attention map in Inter-SIM. As aforementioned, our method is still effective on DFDC dataset which contains many partially forged videos. In verification, we illustrates the U-T attention map, calculated by $\bar{F}_1 \odot \bar{F}_2$ in inter-SIM, to find out whether our model is able to locate the forge frames. As illustrated in Figure 5, our model successfully locates the fake faces on partially forged videos with short and long term movements. Besides, thanks to our snippet-based learning strategy and the elaborately designed Intra- and Inter-SIM modules, even the fake frame in partially forged snippet can also be spotted. This demonstrates the effectiveness of our method.

Conclusion

In this paper, we present a novel DeepFake video detection framework by focusing on the local inconsistency in snippets, which contain a few successive video frames. Built on these snippets, our framework consists of an Intra-Snippet Inconsistency Module (Intra-SIM) for local inconsistency modelling and an Inter-Snippet Interaction Module (Inter-SIM) for cross-snippet interaction promotion. The proposed method outperforms state-of-the-art on four popular benchmarks. In-depth ablation studies and visualizations further demonstrate its effectiveness.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (No. 61972157, No. 72192821), National Key Research and Development Program of China (No. 2019YFC1521104), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200) and Art major project of National Social Science Fund (I8ZD22).

References

Agarwal, S.; and Farid, H. 2021. Detecting Deep-Fake Videos From Aural and Oral Dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 981–989.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; and Ji, R. 2021. Local Relation Learning for Face Forgery Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1081–1088.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, 5781–5790.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 248–255.

Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. C. 2019. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.

Gu, Z.; Chen, Y.; Yao, T.; Ding, S.; Li, J.; Huang, F.; and Ma, L. 2021. Spatiotemporal Inconsistency Learning for DeepFake Video Detection. arXiv:2109.01860.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips Don't Lie: A Generalisable and Robust Approach To Face Forgery Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5039–5049.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hu, Z.; Xie, H.; Wang, Y.; Li, J.; Wang, Z.; and Zhang, Y. 2021. Dynamic Inconsistency-aware DeepFake Video Detection. In *Proceedings of the Thirtieth International Joint Conference on ArtificialIntelligence*, 736–742.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koujan, M. R.; Doukas, M. C.; Roussos, A.; and Zafeiriou, S. 2020. Head2head: Video-based neural head synthesis. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 16–23. IEEE.

Li, J.; Xie, H.; Li, J.; Wang, Z.; and Zhang, Y. 2021. Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6458–6467.

Li, X.; Lang, Y.; Chen, Y.; Mao, X.; He, Y.; Wang, S.; Xue, H.; and Lu, Q. 2020a. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM international conference on multimedia*, 1864–1872.

Li, Y.; and Lyu, S. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 46–52.

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

Lin, J.; Gan, C.; and Han, S. 2019. Temporal shift module for efficient video understanding. 2019 IEEE. In *Proceedings of the IEEE International Conference on Computer Vision*, 7082–7092.

Liu, Z.; Luo, D.; Wang, Y.; Wang, L.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Lu, T. 2020. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11669–11676.

Masi, I.; Killekar, A.; Mascarenhas, R. M.; Gurudatt, S. P.; and AbdAlmageed, W. 2020. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*, 667–684. Springer.

Matern, F.; Riess, C.; and Stamminger, M. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In *Proceedings of the IEEE Winter Applications of Computer Vision Workshops*, 83–92. IEEE.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2307–2311.

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE international conference on computer vision*, 7184–7193.

Qi, H.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Feng, W.; Liu, Y.; and Zhao, J. 2020. DeepRhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4318–4327.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, 86–103. Springer.

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

Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; and Natarajan, P. 2019. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1): 80–87.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sohrawardi, S. J.; Chintha, A.; Thai, B.; Seng, S.; Hickerson, A.; Ptucha, R.; and Wright, M. 2019. Poster: Towards robust open-world detection of deepfakes. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2613–2615.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2014. C3D: generic features for video analysis. *CoRR, abs/1412.0767*, 2(7): 8.

Wang, G.; Zhou, J.; and Wu, Y. 2020. Exposing Deepfaked Videos by Anomalous Co-motion Pattern Detection. In *arXiv preprint arXiv:2008.04848*.

Wang, L.; Tong, Z.; Ji, B.; and Wu, G. 2021. TDN: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1895–1904.

Wang, X.; Yao, T.; Ding, S.; and Ma, L. 2020. Face manipulation detection via auxiliary supervision. In *International Conference on Neural Information Processing*, 313–324. Springer.

Wu, W.; Zhao, Y.; Xu, Y.; Tan, X.; He, D.; Zou, Z.; Ye, J.; Li, Y.; Yao, M.; Dong, Z.; et al. 2021. DSANet: Dynamic Segment Aggregation Network for Video-Level Representation Learning. *arXiv preprint arXiv:2105.12085*.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503.

Zhang, S.; Guo, S.; Huang, W.; Scott, M. R.; and Wang, L. 2020. V4d: 4d convolutional neural networks for video-level representation learning. *arXiv preprint arXiv:2002.07442*.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021. Multi-attentional Deepfake Detection. *CoRR*, abs/2103.02406.

Zi, B.; Chang, M.; Chen, J.; Ma, X.; and Jiang, Y.-G. 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2382–2390.