DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms

Hua Qi[1*], Qing Guo[2*], Felix Juefei-Xu[3], Xiaofei Xie[2], Lei Ma[1†]
Wei Feng[4], Yang Liu[2], Jianjun Zhao[1]
[1]Kyushu University, Japan [2]Nanyang Technological University, Singapore
[3]Alibaba Group, USA [4]Tianjin University, China

ABSTRACT

As the GAN-based face image and video generation techniques, widely known as DeepFakes, have become more and more matured and realistic, there comes a pressing and urgent demand for effective DeepFakes detectors. Motivated by the fact that remote visual photoplethysmography (PPG) is made possible by monitoring the minuscule periodic changes of skin color due to blood pumping through the face, we conjecture that normal heartbeat rhythms found in the real face videos will be disrupted or even entirely broken in a DeepFake video, making it a potentially powerful indicator for DeepFake detection. In this work, we propose *DeepRhythm*, a DeepFake detection technique that exposes DeepFakes by monitoring the heartbeat rhythms. *DeepRhythm* utilizes dual-spatial-temporal attention to adapt to dynamically changing face and fake types. Extensive experiments on FaceForensics++ and DFDC-preview datasets have confirmed our conjecture and demonstrated not only the effectiveness, but also the generalization capability of *DeepRhythm* over different datasets by various DeepFakes generation techniques and multifarious challenging degradations.

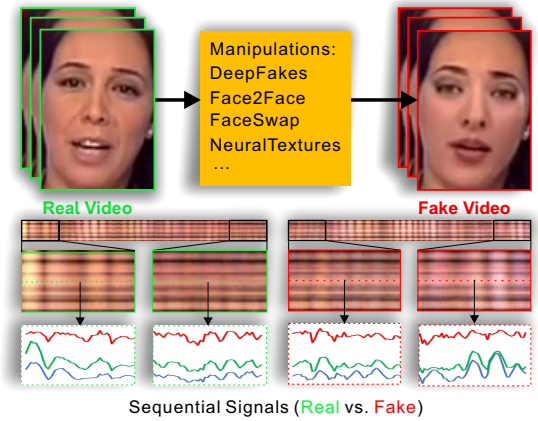


Figure 1: An example of a real video and its fake video generated by various manipulations, *e.g.*, DeepFakes, Face2Face, FaceSwap, *etc.* [57]. It is hard to decide real/fake via the appearance from a single frame. The state-of-the-art Xception [9] fails in this case. However, we see that the manipulations easily diminish the sequential signals representing remote heartbeat rhythms.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Security and privacy** → **Social aspects of security and privacy**.

KEYWORDS

DeepFake detection, heartbeat rhythm, remote photoplethysmography (PPG), dual-spatial-temporal attention, face forensics

ACM Reference Format:
Hua Qi[1*], Qing Guo[2*], Felix Juefei-Xu[3], Xiaofei Xie[2], Lei Ma[1†] and Wei Feng[4], Yang Liu[2], Jianjun Zhao[1]. 2020. *DeepRhythm*: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413707

[*] Both authors contributed equally to this research.
[†] Lei Ma is the corresponding author (malei@ait.kyushu-u.ac.jp).

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413707

1 INTRODUCTION

Over the past decades, multimedia contents such as image and video have become more and more prevalent on various social media platforms. More recently, with the advancement in deep learning-based image and video generation techniques, *i.e.*, generative adversarial networks (GAN) [22], anyone can now generate, *e.g.*, a realistic-looking face that does not exist in the world, or perform a face swap in a video with a high level of realism. The latter is what the community refers to as the DeepFake [36–38]. Such a face swap used to require domain expertise such as theatrical visual effects (VFX) and/or high-speed tracking with markers (*e.g.*, motion captures in the movie Avatar). But now, anyone can do it easily. The low barriers to entry and wide accessibility of pre-trained DeepFake generator models are what the problem is. DeepFakes are now a pressing and tangible threat to the integrity of multimedia information available to us. DeepFakes, *e.g.*, when applied on politicians, fueled with targeted misinformation, can really sway people's opinions and can lead to detrimental outcomes such as manipulated and interfered election without people even knowing about it.

Therefore, the fight against DeepFakes is currently in dire need. Although the detection of DeepFakes is a fairly new research frontier, there have been some attempts to spot DeepFake videos. Some methods are based on traditional digital forensics techniques (see Section 2.2), while others heavily rely on deep learning-based image and video classification *(real vs. fake)* from the raw pixel-domain DeepFake inputs. However, such detection methods based solely on the raw pixel-domain input might become less effective when the DeepFake images and videos become more and more realistic as the deep image generation methods themselves become more mature in the near future. Consequently, a fundamentally different DeepFake detection method is needed.

In this work, we present *DeepRhythm*, a novel DeepFake detection technique that is intuitively motivated and is designed from ground up with first principles in mind. Motivated by the fact that remote visual photoplethysmography (PPG) [69] is made possible by monitoring the minuscule periodic changes of skin color due to blood pumping through the face from a video, we conjecture that normal heartbeat rhythms found in the real face videos will be disrupted or even broken entirely in a DeepFake video, making it a powerful indicator for detecting DeepFakes. As shown in Figure 1, existing manipulations, *e.g.*, DeepFakes, significantly change the sequential signals of the real video, which contains the primary information of the heartbeat rhythm. To further make our proposed DeepRhythm method work more robustly under various degradations, we have devised and incorporated both a heart rhythm motion amplification module as well as a learnable spatial-temporal attention mechanism at various stages of the network model.

Together, through extensive experiments, we have demonstrated that our conjecture holds true and the proposed method indeed effectively exposes DeepFakes by monitoring the heartbeat rhythms. More specifically, DeepRhythm outperforms four state-of-the-art DeepFake detection methods including Bayer's method [3], Inception ResNet V1 [59], Xception [9], and MesoNet [1] in the FaceForensics++ benchmark [57] and exhibits high robustness to JPEG compression, noise, and blur degradations. To the best of our knowledge, this is the very first attempt to expose DeepFakes using heartbeat rhythms. Our main contributions are summarized as follows:

- We propose *DeepRhythm*, the very first method for effective detection of DeepFake with the heartbeat rhythms.
- To characterize the sequential signals of face videos, we propose the *motion-magnified spatial-temporal representation (MMSTR)* that provides powerful discriminative features for high accurate DeepFake detection.
- To fully utilize the MMSTR, we propose *dual-spatial-temporal attention network* to adapt to dynamically changing faces and various fake types. Experimental results on FaceForensics++ and DeepFake Detection Challenge-preview dataset demonstrate that our method not only outperforms state-of-the-art methods but is robust to various degradations.

2 RELATED WORK

2.1 DeepFakes Generation

Recently, DeepFake techniques have gained widespread attention and been used in generating pornographic videos, fake news, and hoaxes, *etc.* Some early studies use face-wrap-based methods to generate fake videos. For example, Bregler *et al.* [4] track the movement of the speaker's mouth and morph the input video. Dale *et al.* [11] present a video face replacement via a face 3D model. Similarly, Garrido *et al.* propose a face warp system while keeping the original face performance [18] and a photo-realistically replacement method via high-quality monocular capture [19]. Thies *et al.* develop a real-time expression transfer for facial reenactment [61] and propose the Face2Face method [62] that tracks target and source facial expressions to build a face 3D model and re-renders source face on the target model. In addition, Thies *et al.* [60] further use neural textures and defer neural render to generate the forgeries.

Besides the above face-wrap-based methods, recent DeepFake approaches, *i.e.*, PGGAN [36], StyleGAN [37], and StyleGAN2 [38], employ the generative adversarial network (GAN) [22] for the near-realistic face synthesis. Moreover, some methods can even alter face attributes, *e.g.*, changing or removing the color of the hair, adding glasses or scars [8, 21, 30, 63], and modifying persons' facial expression [41]. Overall, GANs have shown great potential in this area and are easy to use. However, current DeepFake methods, even those based on GANs, do not explicitly preserve the pulse signal, inspiring us to capitalize on the pulse signal to distinguish the real and manipulated videos.

2.2 Forgery and DeepFake Detection

DeepFake detection is challenging since GAN-based DeepFakes can generate near-realistic faces that are hardly detected even using the state-of-the-art digital forensics. To alleviate this challenge, researchers are exploring effective solutions to identify fake videos.

Early attempts focus on detecting forgeries via hand-crafted features, *e.g.*, [5, 17, 20, 53]. However, these hand-crafted models can be strenuous due to the realistic faces generated by SOTA DeepFake methods (*e.g.*, FaceApp, Reflect, and ZAO). Later, researchers regard the DeepFake detection as a classification problem by extracting discriminative features, *e.g.*, color cues [46], monitoring neuron behaviors [43–45, 66, 68], and employing classifiers, *e.g.*, support vector machine (SVM), to tell whether a video is fake or real.

In addition, many researchers also employ SOTA deep neural networks (DNNs) to detect forgery images. Cozzolino *et al.* [10] use residual-based local features and achieve significant performance. Bayar *et al.* [3] and Rahmouni *et al.* [56] propose novel DNNs to detect manipulated images. Zhou *et al.* [71] combine a DNN-based face classification stream with a steganalysis-based triplet stream, yielding good performance. More recently, researchers are trying to apply much more complex and advanced DNNs on video forgery detection, such as Inception-ResNet [59], MesoNet [1], capsule networks [47], and Xception [9].

Besides only adopting convolutional neural networks (CNNs), some researchers use a combined recurrent neural network (RNN) and CNN to extract image and temporal features to distinguish real and fake videos. For example, Güera *et al.* [28] use CNN features to train an RNN to classify videos. Similarly, Sabir *et al.* [58] use a sequence of spatio-temporal faces, as RNN's input to classify the videos. Furthermore, Dong *et al.* [12] utilize an attention mechanism to generate and improve the feature maps, which highlight the informative regions. Different from existing methods, our technique initiates the first step of leveraging remote heartbeat rhythms

for DeepFake detection. To achieve high detection accuracy, we propose a motion-magnified representation for the heartbeat rhythms and employ a spatial-temporal representation to improve the ability in distinguishing real and fake videos.

2.3 Remote Photoplethysmography (rPPG) Anti-Spoofing

Face spoof detection is similar to DeepFake detection, aiming to determine whether a video contains a live face. Since the remote heart rhythm (HR) measuring techniques achieve quite a bit of progress [2, 6, 35, 39, 54, 55, 69], many works use rPPG for face spoofing detection. For example, Li *et al.* [40] use the pulse difference between real and printed faces to defend spoofing attacks. Nowara *et al.* [49] compare PPGs of face and background to decide whether the face is live or not. Heusch *et al.* [32] use the long-term statistical spectral on the pulse signals and Hernandez-Ortega *et al.* [31] employ the near infrared against realistic artifacts. Moreover, combining with DNNs, Liu *et al.* [42] extract spatial and temporal auxiliary information, *e.g.*, depth map and rPPG signal, to distinguish whether it is a live face or spoofing face.

Overall, existing anti-spoofing methods also benefit from employing rPPG for liveness detection, which seems similar to our work. However, there are fundamental differences: the liveness detection mainly relies on the judgment about whether the heart rhythms exist or not; our work aims to find the different patterns between real and fake heart rhythms since fake videos may still have the heart rhythms but their patterns are diminished by DeepFake methods and are different from the real ones (see Figure 1).

3 METHOD

We propose *DeepRhythm* (Sec. 3.1) for effective DeepFake detection by judging whether the normal HR in face videos are diminished. Figure 2 (a) summarizes the workflow of DeepRhythm.

3.1 DeepRhythm for DeepFake Detection

Given a face video $\mathcal{V} = \{\mathbf{I}_i\}_{i=1}^{T}$ that contains T frames, our goal is to predict if this video is real or fake according to the heart rhythm signals. To this end, we first develop the *motion-magnified spatial-temporal representation (MMSTR) (Sec. 3.2)* for face videos, which can highlight the heart rhythm signals and output a *motion-magnified spatial-temporal map (MMST map)*, *i.e.*, $\mathbf{X} = \text{mmstr}(\mathcal{V}) \in \mathbb{R}^{T\times N\times C}$ where T is the number of frames, N is the N region of interest (ROI) blocks of the face in $\mathcal{V}$ (*i.e.*, the regions marked by the blur grid in Figure 2 (a)), and C means the number of color channels. In the following, we formulate with single color channel for clear representation but use RGB channels in practice. Intuitively, $\mathbf{X}$ contains the motion-magnified temporal variation of N blocks in the face video, *i.e.*, highlighted heart rhythm signals.

We can simply design a deep neural network that takes the MMST map as input and predict if the raw video is real. However, various interference, *e.g.*, head movement, illumination variation, and sensor noises, may corrupt the MMST map. As a result, the contributions of different positions in the MMST map are not always the same (*e.g.*, the three patches shown in Figure 2 (a) have different heart rhythm strength), which definitely affects the fake detection accuracy. To alleviate this challenge, we should assign different weights to different positions of the MMST map before further performing the fake detection

$$\mathbf{y} = \phi(\mathbf{A} \odot \mathbf{X}), \tag{1}$$

where $\phi(\cdot)$ is a CNN for real/fake classification, $\odot$ denotes the element-wise multiplication, and $\mathbf{y}$ is the prediction (*i.e.*, 1 for fake and 0 for real). The matrix $\mathbf{A} \in \mathbb{R}^{T\times N}$ provides different weights to different positions of $\mathbf{X}$ and is known as an attention mechanism. We let RGB channels share the attention matrix.

We aim to produce $\mathbf{A}$ via a DNN. However, due to the diverse types of fake and dynamic changing faces, it is difficult to get proper $\mathbf{A}$ for different face directly. We handle this problem by further decomposing $\mathbf{A}$ into two parts, *i.e.*, spatial attention $\mathbf{s} \in \mathbb{R}^{N\times 1}$ and temporal attention $\mathbf{t} \in \mathbb{R}^{T\times 1}$, and reformulate Eq. (1) as

$$\mathbf{y} = \phi((\mathbf{t} \cdot \mathbf{s}^\top) \odot \mathbf{X}), \tag{2}$$

Intuitively, the two attentions indicate when (along the T's axis) and where (along the N's axis) of the input MMST map should be used for better fake detection. Furthermore, the number of parameters of $\mathbf{s}$ and $\mathbf{t}$, *i.e.*, $N + T$, is much smaller than that of $\mathbf{A}$, *i.e.*, $N \cdot T$, which allows the spatial-temporal-attention to be tuned more easily.

Then, the key problem is how to generate $\mathbf{t}$ and $\mathbf{s}$ to adapt to dynamically changing faces and various fake types. In Sec. 3.3, we propose the dual-spatial-temporal attention network to realize Eq. (2) by jointly considering prior & adaptive spatial attention and frame & block temporal attention.

3.2 Motion-Magnified Spatial-Temporal Representation

A straightforward way of employing heart rate (HR) signals for DeepFake detection is to use existing HR representations that are designed for the remote HR estimation. For example, we can use the spatial-temporal representation (STR) proposed by Niu *et al.* [48] for representing HR signals and feed them to a classifier for DeepFake detection. However, it is hard to achieve high fake detection accuracy with the STR directly since the differences between real and fake videos are not highlighted, *i.e.*, STR's discriminative power for DeepFake detection is limited.

To alleviate the problem, we propose the *motion-magnified STR (MMSTR)* where differences between real and fake face videos can be effectively represented. Specifically, given in a face video, *i.e.*, $\mathcal{V}$ having T frames, we calculate MMSTR using the following steps:

(i) Calculate landmarks[1] of the faces on all frames of $\mathcal{V}$ and remove the eyes and background according to the landmarks, *e.g.*, the faces shown in the left of Figure 2 (b).

(ii) Perform the motion magnification algorithm [50, 67][2] on the background removed face video and obtain motion-magnified face video with RGB space.

(iii) Divide the face areas of all frames into N non-overlapping ROI blocks, *i.e.*, regions marked by the blue grid in Figure 2 (b), and perform average pooling on each block and each color channel for each frame. We then obtain the MMST map, *i.e.*, $\mathbf{X}$, as the sub-figures shown in the right of Figure 2 (b). Each row of $\mathbf{X}$ represents the motion-magnified

[1] https://github.com/codeniko/shape_predictor_81_face_landmarks.
[2] We use its python implementation: https://github.com/flyingzhao/PyEVM.

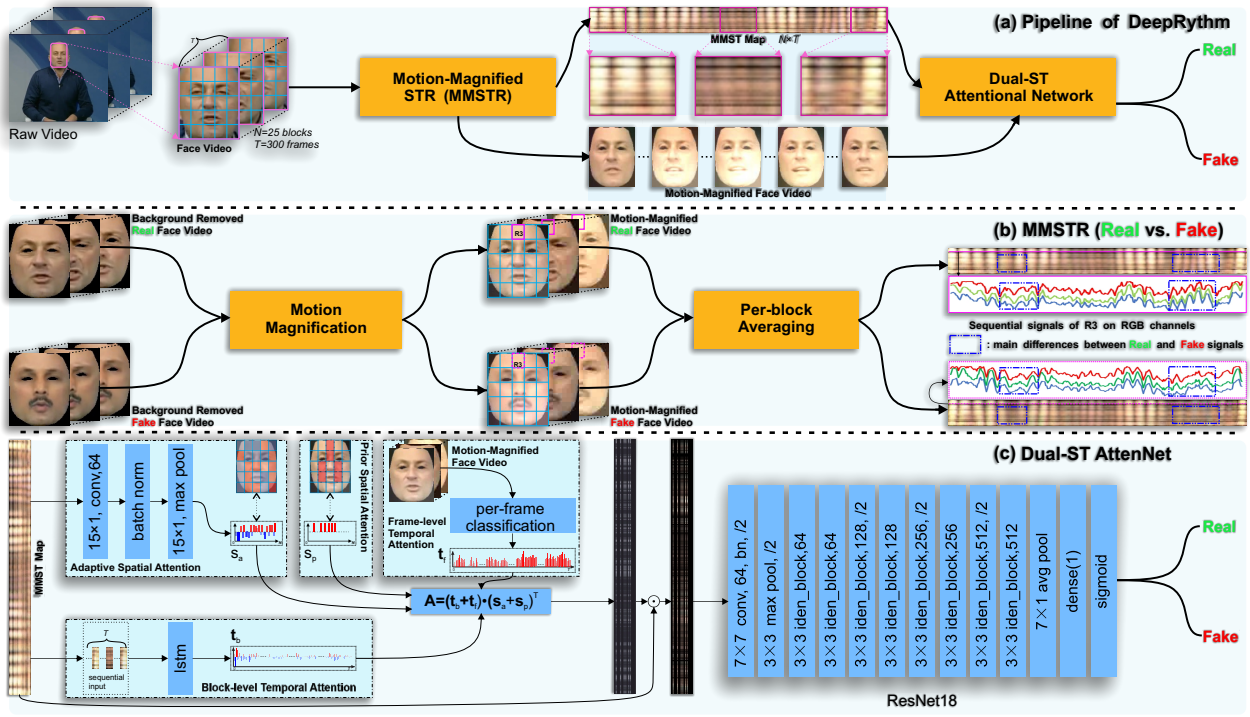


Figure 2: The workflow of DeepRhythm, *i.e.*, (a), and its two main modules: motion-magnified spatial-temporal representation (MMSTR), *i.e.*, (b), and dual-spatial-temporal attentional network (Dual-ST AttenNet), *i.e.*, (c). We also highlight the main differences of MMSTRs between real and fake videos in sub-figure (b).

temporal variation of one block on RGB channels, as the red, green, and blue curves shown in Figure 2 (b).

Figure 2 (b) shows examples of real and fake face videos and their MMST maps, respectively. We have the following observations: 1) it is difficult to judge which video is fake just by looking at the raw frames. 2) differences between the real and fake videos can be easily found on our MMST maps that will provide effective information for fake detection. The advantages of MMSTR over STR will be further discussed in the experimental section.

3.3 Dual-Spatial-Temporal Attentional Network

In this section, we detail the dual-spatial-temporal attentional network (Dual-ST AttenNet), with which we can realize accurate DeepFake detection through the MMST map and its spatial and temporal attentions, *i.e.*, $\mathbf{t}$ and $\mathbf{s}$ defined in Eq. (2).

3.3.1 Dual-Spatial Attention. The dual-spatial attention is

$$\mathbf{s} = \mathbf{s}_a + \mathbf{s}_p, \tag{3}$$

where $\mathbf{s}_p \in \mathbb{R}^{N\times 1}$ and $\mathbf{s}_a \in \mathbb{R}^{N\times 1}$ are the prior and face-adaptive spatial attentions, respectively. The prior attention $\mathbf{s}_p$ is a fixed vector whose six specified elements are set as one while others are zero, which is to extract the HR signals from six specified ROI blocks and ignore signals from other blocks. The six specified ROI blocks are the four blocks under eyes and two blocks between eyes, as shown in Figure 2 (c). The intuition behind this idea is that the specified blocks are usually robust to various real-world interference while the HR signals of other blocks are easily diminished when unexpected situations happen, *e.g.*, head movement might let the HR signals of blocks at face boundary disappear.

In addition to the prior spatial attention, we also need face-adaptive attention, *i.e.*, $\mathbf{s}_a$, to highlight different blocks to adapt to the environment variations since even the same face under different situations, *e.g.*, the illumination changes, has different effective ROI blocks. To this end, we propose to train a spatial attention network to generate adaptive spatial attention, which contains a convolution layer that has 64 kernels with size being 15×1 followed by a batch normalization layer and max-pooling layer. The CNN's parameters are jointly learned with the whole framework.

3.3.2 Dual-Temporal Attention. DeepFake methods usually add different fake textures at different face locations to different frames, which not only destroy the smooth temporal variation of a face but lead to inconsistent fake magnitude among frames (*i.e.*, some frames contain obvious fake textures while others have few or no fakes). We propose dual-temporal attention to consider above information

$$\mathbf{t} = \mathbf{t}_b + \mathbf{t}_f, \tag{4}$$

where $\mathbf{t}_b \in \mathbb{R}^{T\times 1}$ and $\mathbf{t}_f \in \mathbb{R}^{T\times 1}$ indicate which frames are more significant for final fake detection. Specifically, we train an LSTM to represent the temporal variation of a face, which is sequentially fed with each row of the MMST map, *i.e.*, $\mathbf{X}$, and outputs $\mathbf{t}_b$ that is denoted as the block-level temporal attention. The LSTM's parameters are jointly trained with the whole framework.

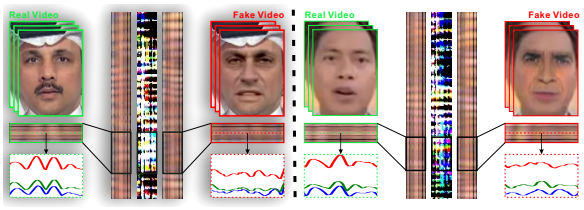


Figure 3: Two real-fake video pairs, their MMST maps and the colorful difference maps between real and fake.

To take full advantage of fake textures in each frame, we train a temporal-attention network that takes each motion-magnified frame as input and scores the fakeness of the frame independently and get $\mathbf{t}_f$. The frames with higher probability to be fake contribute more to the final classification and we denote $\mathbf{t}_f$ as the frame-level temporal attention. In practice, we use the Meso-4 architecture [1] as the network containing a sequence of four convolution layers and two fully-connected layers. The Meso-4's parameters are independently trained for frame-level fake detection.

3.3.3 Implementation details. Our dual-spatial-temporal attentional network is shown in Figure 2 (c) where an MMST map, *i.e.*, $\mathbf{X}$, is first employed to produce the adaptive spatial attention, *i.e.*, $\mathbf{s}_a$ and the block-level temporal attention, *i.e.*, $\mathbf{t}_b$, through a spatial attention network and an LSTM, respectively. The pre-trained Meso-4 is fed with the motion-magnified face video and outputs the frame-level temporal attention, *i.e.*, $\mathbf{t}_f$. Finally, the attentional MMST map, *i.e.*, $(\mathbf{t} \cdot \mathbf{s}^\top) \cdot \mathbf{X}$, is fed to the $\phi(\cdot)$ for the final DeepFake detection where we use ResNet18 [29] for the $\phi(\cdot)$.

We jointly train parameters of the spatial attention network, the LSTM, and the ResNet18 using the cross-entropy loss with Adam optimizer. The learning rate and weight decay are set as 0.1 and 0.01 respectively. The max epoch number is set to 500, and training will stop if validation loss did not decrease in 50 epochs. For training the Meso-4, we use the same hyper-parameters. We use videos from FaceForensics++ [57] as the training dataset which is introduced in Sec. 4.1 and Table 1. Our implementation and results are obtained on a server with Intel Xeon E5-1650-v4 CPU and NVIDIA GP102L.

4 EXPERIMENTS

4.1 Dataset and Experiment Setting

Dataset. We select FaceForensics++ [57] as our training and testing datasets, and use DeepFake Detection Challenge preview (DFDC-preview) [15] as an additional testing dataset to evaluate our method's cross-dataset generalization capability.

FaceForensics++ dataset consists of thousands of videos manipulated with different DeepFake methods and contains four fake sub-datasets, *i.e.*, DeepFake Detection (DFD), DeepFake (DF), Face2Face (F2F), and FaceSwap (FS). However, the original FaceForensics++ dataset has the data imbalance problem. For example, the original DFD subset contains 2728 fake videos, but only 268 real videos. To solve this problem, we make the following improvements: 1) we augment the the original 268 real videos by flipping them horizontally and get a total of 2,510 real videos. 2) To evaluate our method on the whole FaceForensics++ dataset, we build an extra dataset, *i.e.* 'ALL' in Table 1, by concatenating the four subsets and augment the real videos by flipping them horizontally and vertically, and rotating them 180 degrees, respectively. Table 1 summarizes the final four subsets, the ALL dataset, and their partitions about training, validation, and testing. The dataset partition ratio is 8:1:1 and the augmented videos are removed in the testing datasets.

Table 1: Details of FaceForensics++ (FF++) and DFDC-preview datasets for both testing and training

Dataset		total	real	fake	train	val.	test
FF++	DFD	5238	2510	2728	4190	524	524
	DF	1959	988	971	1567	195	197
	F2F	1966	988	978	1572	196	198
	FS	1971	988	983	1582	197	198
	ALL	10680	5020	5660	8544	1068	1068
DFDC-preview		3310	578	2732	-	-	1000

For videos in DeepFake Detection Challenge, we directly use it as the testing set. The details can be found in Table 1. Directly using the DFDC-preview dataset as testing set will cause imbalance between the number of real and fake videos. So we randomly sample 500 videos from the real part and 500 videos from the fake part, and assemble them as the testing set. Then, we test every method formerly trained on ALL's training set on it to compare their performance.

Pre-processing. For every video in the testing and training datasets, we take the first 300 frames to produce the MMST map. Specifically, while processing frames, we first use MTCNN [52] to detect face, and then use Dlib to get 81 facial landmarks [51]. If faces are not detected in a frame, this frame will be abandoned. If more than 50 frames were abandoned, this video will not be used to train the network. If more than one faces were detected in one frame, the one closer to the faces of previous frames will be retained.

Baseline. We choose the state-of-the-art DeepFake detection methods, *i.e.*, Bayer's method [3], Inception ResNet V1 [59], Xception [9] and MesoNet [1], as baselines. All of them obtain high performance on FaceForensics++'s benchmark [13] and are the top-4 accessible methods therewithin. For Bayer's method [3], we do not find publicly available code, so we re-implement it on Keras. For Inception ResNet V1 [59] and Xception [9], we directly use the Keras code provided and only add a Dense layer with one neuron after the final layer to get prediction. As for MesoNet [1], we directly use the code provided by authors [14].

It should be noted that these baselines perform the fake detection on an image instead of a video, *i.e.*, estimating if a frame is real or fake. We make the following adaptations to make them suitable to address videos: 1) For the testing setup, we use these baselines to predict every frame of the video and count the number of real or fake frames. If the real frames are more than fake ones, we identified this video as real, and vice versa. 2) In terms of the training setup, we take the first five frames for every training video in the 'ALL' dataset in Table 1 and extract their facial region via MTCNN, all of these faces are divided into training, validation, and testing subsets. We also employ Adam optimizer with batch-size of 32 and learning rate of 0.001. The max epoch number is 500, and the training will stop if validation loss did not decrease in 50 epochs.

Table 2: Comparison with baseline methods on FaceForensics++ and DFDC-preview datasets with the models trained on sub-datasets and ALL dataset of FaceForensics++, respectively. We highlight the best and second best results with red and yellow.

	train on sub-datasets				train on ALL dataset					
test on	DFD	DF	F2F	FS	DFD	DF	F2F	FS	ALL	DFDC
Bayer and Stamm [3]	0.52	0.503	0.505	0.505	0.501	0.52	0.503	0.505	0.5	0.5
Inception ResNet V1 [59]	0.794	0.783	0.788	0.778	0.919	0.638	0.566	0.462	0.774	0.597
Xception [9]	0.98	0.995	0.985	0.98	0.965	0.984	0.984	0.97	0.978	0.612
MesoNet [1]	0.804	0.979	0.985	0.995	0.958	0.822	0.813	0.783	0.909	0.745
DeepRhythm (ours)	0.987	1.0	0.995	1.0	0.975	0.997	0.989	0.978	0.98	0.641

4.2 Baseline Comparison on Accuracy

We test all methods on DFD, DF, F2F, FS, and ALL subsets (reported in Table 2) and the DFDC-preview with the models trained on FF++'s subsets (DFD, DF, F2F, FS, and ALL), respectively.

Results on FaceForensics++. Overall, in Table 2, our DeepRhythm achieves the highest accuracy on all datasets compared with the baseline methods. First, our method gets better results than other methods across all cases, regardless of which training dataset is used. This demonstrates the generalization capability of our method across various DeepFake techniques. Second, although we adopt the MesoNet in our framework for the frame-level temporal attention, our method significantly outperforms the MesoNet on all cases, *e.g.*, when trained on ALL dataset, DeepRhythm achieves 0.96 on the FS while MesoNet only has 0.719, which validates the effectiveness of our MMST representation and other attention information, and also indicates the potential capability of our framework for enhancing existing frame-level DeepFake detection methods. Third, although the baseline method Xception has obtained significantly high accuracy, *e.g.*, 0.985 and 0.995 on the testing dataset of F2F and FS, it is still exceeded by our method, confirming the advantage of our method over the state-of-the-arts. We show two cases from the FaceForensics++ dataset in Figure 3, where all baseline methods fail to recognize the fake videos while our method succeeds. The fake techniques diminish the sequential signal patterns of real videos (*e.g.*, the waveform of the real video in the first case becomes flat in the fake video), which are effectively captured by our MMST maps.

Results on DFDC-preview. According to the results on the DFDC (see Table 2), Bayer's still performs worse than others, achieving 0.5 accuracy. Inception ResNet V1 has worse performance than their results on ALL's testing set, achieving 0.597. Xception obtains 0.612 accuracy on DFDC-preview, and is also much worse than its performance on ALL's testing set. Our DeepRhythm gets 0.641 accuracy and is better than Xception, being the second highest accuracy. Although MesoNet performs not as good on ALL's testing set, it achieves the highest accuracy (*i.e.*, 0.745).

4.3 Ablation Study on Accuracy

To demonstrate the effectiveness of our motion-magnified spatial-temporal representation (MMSTR), dual-spatial-temporal attention network, and end-to-end training, we conduct an ablation study by first training the basic model with existing spatial-temporal (ST) map at the beginning and then add our contributions one by one.

DeepRhythm variants. We first train the bare model (DR-st), which only uses ST map as its input without motion magnification

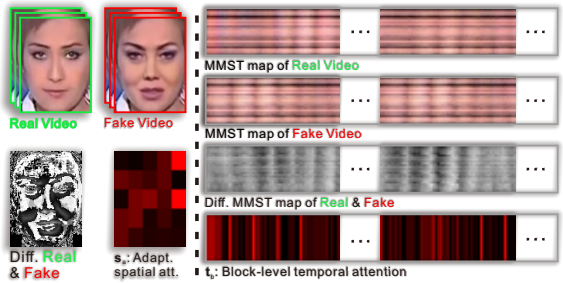


Figure 4: An example of a real video, the corresponding fake video, the difference image between real and fake frames (Diff. Real & Fake), the MMST maps of real and fake videos, the difference map between real and fake MMST maps (Diff. MMST map of Real & Fake), the adaptive spatial attention (s_a, *i.e.*, adapt. spatial att.), and the block-level temporal attention (t_b).

and attention. Then, we use *MMST map* as inputs and re-train our model (denoted as DR-mmst), still not using any attention. After that, based on the pre-trained DR-mmst, we add *adaptive spatial attention* (A) and *block-level temporal attention* (B), respectively, (*i.e.*, DR-mmst-A and DR-mmst-B) and perform fine-tuning. After 50 epochs, we observe that the validation loss does not decrease further. Then, we add *prior spatial attention* (P) and *frame-level temporal attention* (F) on DR-mmst-A and DR-mmst-B, respectively, then get DR-mmst-AP and DR-mmst-BF that are further fine-tuned. Next, based on either DR-mmst-AP or DR-mmst-BF, we use four attentions together (*i.e.*, DR-mmst-APBF) and carry on training. Finally, we compare DR-mmst-APBF with our final version (*i.e.*, DR-mmst-APBF-e2e), where the *adaptive spatial attention* (A), *block-level temporal attention* (B), and the network are jointly or end-to-end trained. For all the experiments, we use the same hyper-parameters and datasets, as introduced in Sec. 3.3.3. The results are summarized in Table 3.

Effectiveness of MMSTR. As shown in Table 3, our MMSTR significantly improves the DR-st's accuracy, *e.g.*, 0.328 improvement on ALL. The ST map from [48] has little discriminative power for DeepFake detection since DR-st achieves about 0.5 accuracy on every testing dataset, which means it randomly guesses a video being real/fake. After using our MMSTR, DR-mmst achieves 0.217 averaged accuracy increment over DFD, DF, F2F, FS, and ALL datasets.

Effectiveness of single attention. Based on the DR-mmst, we add adaptive spatial attention (DR-mmst-A) and block-level temporal attention (DR-mmst-B), respectively. These two attentions do help improve the model's accuracy, as presented in Table 3

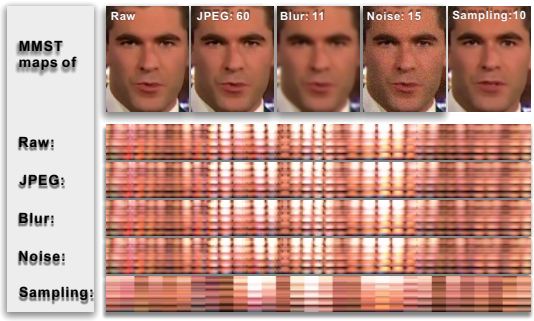


Figure 5: Frames and MMST maps of a video and their four degraded versions with JPEG, blur, noise, and temporal sampling degradations whose degrees are 60, 11, 15, and 10, respectively, which are the median values in the x-axis ranges in Figure 7. Clearly, the JPEG, blur, and noise degradations do not affect the MMST maps of raw videos. The temporal sampling significantly diminishes the raw pattern of the MMST maps.

where DR-mmst-A and DR-mmst-B get average 0.061 and 0.0632 improvements over DR-mmst, respectively.

We further show an example of the adaptive spatial attention ($\mathbf{s}_a$) and the block-level temporal attention ($\mathbf{t}_b$) in Figure 4. To validate their effectiveness, we also present the difference image and MMST map between real and fake cases. From the view of spatial domain, the difference image indicates that the main changes caused by the fake is around the nose, which is identical to the estimated adaptive spatial attention. In terms of the temporal domain, the estimated temporal attention has high values at the peaks of the difference MMST map.

Effectiveness of dual-spatial attention. In addition to the adaptive spatial attention (DR-mmst-A), we further consider the prior attention where the specified ROI blocks on faces are considered and realize the DR-mmst-AP. As validated in Table 3, DR-mmst-AP outperforms DR-mmst-A on all compared datasets and obtains an average of 0.033 improvement, which demonstrates the advantage of dual-spatial attention over single adaptive spatial attention.

Effectiveness of dual-temporal attention. The block-level temporal attention misses details among frames. To alleviate this issue, we add the frame-level temporal attention (F) to DR-mmst-B for the frame-level DeepFake detection and get the DR-mmst-BF. In Table 3, DR-mmst-BF has much higher accuracy than DR-mmst-B on all compared datasets. The average improvement is 0.178, which shows the effectiveness of our dual-temporal attention.

Effectiveness of dual-spatial-temporal attention and end-to-end training. We put DR-mmst-AP and DR-mmst-BF together and get DR-mmst-APBF. Compared with DR-mmst-AP, DR-mmst-APBF has much higher accuracy on all datasets. However, when comparing it with DR-mmst-BF, DR-mmst-APBF's accuracy slightly decreases on DFD, DF, and ALL while increasing on FS dataset. Though, when we train DR-mmst-APBF in the end-to-end way and get DR-mmst-SPTM-e2e, it achieves the highest accuracy on all testing datasets, indicating that training four attention separately might not mine the potential power of the four attention effectively, and training them together helps get the maximum effect.

Table 3: Ablation study of DeepRhythm (DR) by progressively adding the MMSTR, adaptive (A) and prior (P) spatial attentions, block-level (B) and frame-level (F) temporal attentions, and end-to-end (e2e) training strategy.

	train on ALL sub-dataset				
test on	DFD	DF	F2F	FS	ALL
DR-st	0.522	0.497	0.497	0.492	0.512
DR-mmst	0.814	0.684	0.635	0.64	0.84
DR-mmst-A	0.849	0.77	0.736	0.716	0.847
DR-mmst-B	0.872	0.745	0.731	0.731	0.85
DR-mmst-AP	0.879	0.816	0.766	0.756	0.867
DR-mmst-BF	0.97	0.969	0.954	0.959	0.966
DR-mmst-APBF	0.965	0.959	0.954	0.965	0.964
DR-mmst-APBF-e2e	0.972	0.98	0.964	0.959	0.98

4.4 Baseline Comparison on Robustness

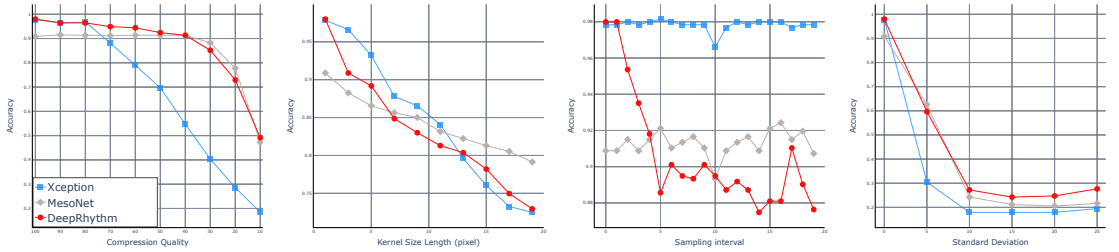
In this section, we study the robustness of our method and two baseline methods, *i.e.*, Xception and MesoNet, which have the highest accuracy among baselines. Their models are trained on the training set of the ALL dataset. We consider four general degradations, *i.e.*, JPEG compression, Gaussian blur, Gaussian noise, and temporal sampling, and construct a degradation dataset by manipulating the testing set of the ALL dataset. For the first three degradations, we add the corresponding interference to each frame of the tested video and use the compression quality, blur kernel size, and standard deviation of noise to control the degradation degree, respectively. We show the degradation examples in Figure 5. The temporal sampling means that we do not use the raw continuous frames to get the MMST map but select frame at every K frames. We use temporal sampling to test if our method still works under the unsmooth temporal variation. Please refer to the x-axis in Figure 6 and 7 for the variation range of each degradation.

As shown in Figure 6, our method exhibits strong robustness on JPEG compression and Gaussian noise, but do not perform well on temporal sampling when compared with Xception and MesoNet. However, we could mitigate the issue with the video frame interpolation techniques, which is yet to be explored as the future work.

4.5 Ablation Study on Robustness

We use the degradation dataset in Sec. 4.4 to analyze the robustness of MesoNet (*i.e.*, the frame-level temporal attention) and seven DeepRhythm variants (see the legend of Figure 7). These methods can be roughly divided into two clusters, one using MesoNet for the frame-level temporal attention (donated as F-cluster), including MesoNet, DR-mmst-BF, DR-mmst-APBF, and DR-mmst-APBF-e2e; the others do not employ the MesoNet (donated as non-F-cluster), including DR-mmst, DR-mmst-A, DR-mmst-B, and DR-mmst-AP.

As shown in Figure 7, our MMSTR helps the variants, *i.e.*, DR-mmst, DR-mmst-A, DR-mmst-B, and DR-mmst-AP, to keep at almost the same accuracy across all compression quality. The reason is that the MMSTR is calculated by average pooling pixel values in ROI blocks, thus is insensitive to local pixel variation caused by JPEG compression, Gaussian blur, and Gaussian noise. As shown in Figure 5, the MMST maps of JPEG compressed, noisy, and blurred videos are almost the same to the raw video. On the other hand, the MesoNet handles frames independently and relies on detailed



Xception	MesoNet	DeepRhythm
0.67±0.283	0.852±0.133	0.871±0.145

(a) JPEG

Xception	MesoNet	DeepRhythm
0.847±0.088	0.843±0.035	0.834±0.072

(b) Blur

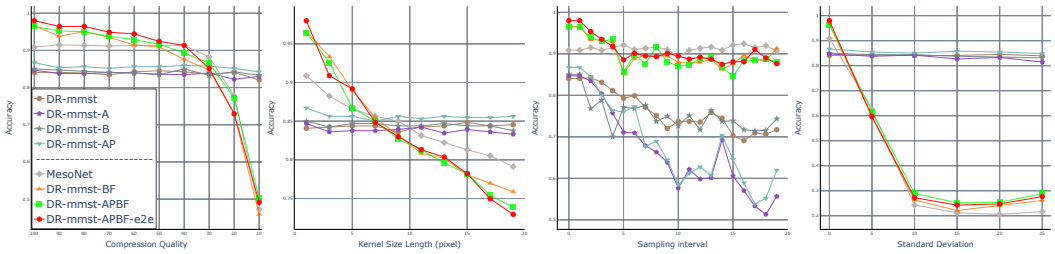
Xception	MesoNet	DeepRhythm
0.847±0.088	0.843±0.035	0.834±0.072

(c) Sampling

Xception	MesoNet	DeepRhythm
0.847±0.088	0.843±0.035	0.834±0.072

(d) Gaussian Noise

Figure 6: Baseline comparison on robustness. We perform DeepFake detection through DeepRhythm and two state-of-the-art baselines, *i.e.*, Xception and MesoNet, on a degradation dataset. Four degradations, *i.e.*, JPEG compression, Gaussian blur, temporal sampling, and Gaussian noise, are added to the testing set of the ALL dataset. The average accuracy and corresponding standard deviation across all degradation degrees are presented at the bottom of each sub-figure.



(a) JPEG (b) Blur (c) Sampling (d) Gaussian Noise

	DR-mmst	DR-mmst-A	DR-mmst-B	DR-mmst-AP	MesoNet	DR-mmst-BF	DR-mmst-APBF	DR-mmst-APBF-e2e
JPEG Compression	0.838±0.007	0.836±0.006	0.841±0.006	0.855±0.006	0.852±0.133	0.853±0.146	0.868±0.133	0.871±0.145
Blur	0.843±0.003	0.838±0.004	0.845±0.003	0.856±0.004	0.843±0.035	0.84±0.067	0.832±0.068	0.834±0.072
Sampling	0.762±0.05	0.668±0.103	0.754±0.04	0.688±0.101	0.913±0.007	0.9±0.03	0.897±0.033	0.906±0.031
Gaussian Noise	0.842±0.002	0.834±0.011	0.84±0.006	0.855±0.006	0.402±0.271	0.425±0.274	0.444±0.265	0.436±0.273

(e) Average accuracy ± standard deviation across degradation degrees

Figure 7: Ablation Study on robustness. We perform DeepFake detection through MesoNet and DeepRhythm's seven variants on a degradation dataset. The compared methods are clustered to two types, *i.e.*, F-cluster using MesoNet for frame-level temporal attention (*i.e.*, MesoNet itself, DR-mmst-BF, DR-mmst-APBF, and DR-mmst-APBF-e2e), and non-F-cluster that does not employ MesoNet (*i.e.*, DR-mmst, DR-mmst-A, DR-mmst-B, and DR-mmst-AP). For each degradation, the average accuracy and corresponding standard deviation across all degradation degrees are presented at the bottom of figure.

information within frames. As a result, it helps our methods be robust to temporal sampling and achieve the best performance but is sensitive to local pixel variation. Clearly, the advantages and disadvantages of MMSTR and MesoNet are complementary. Our final version combining these two modules shows comprehensive robustness across all degradations.

5 CONCLUSIONS

In this work, we have proposed DeepRhythm, a novel DeepFake detection technique. It is intuitively motivated by the fact that remote visual photoplethysmography (PPG) is made possible by monitoring the minuscule periodic changes of skin color due to blood pumping through the face. Our extensive experiments on FaceForensics++ and DFDC-preview datasets confirm our conjecture that normal heartbeat rhythms in the real face videos are disrupted in a DeepFake video, and further demonstrate not only the effectiveness of DeepRhythm, but how it generalizes over different datasets by various DeepFake generation techniques. One interesting future direction is to study the combined effort of DeepRhythm with other DeepFake detectors [33, 34, 65, 66]. Beyond DeepFake detection, the investigation of how DeepRhythm can be applied further to domains such as countering non-traditional adversarial attacks [7, 26, 27, 64] is also potentially viable. In addition, the possibility of using tracking methods [16, 23–25, 70] to mine more discriminative spatial-temporal features would also be further studied.

ACKNOWLEDGMENTS

This research was supported by JSPS KAKENHI Grant No. 20H04168, 19K24348, 19H04086, JST-Mirai Program Grant No. JPMJMI18BB, Japan. It was also supported by Singapore National Cybersecurity R&D Program No. NRF2018NCR-NCR005-0001, National Satellite of Excellence in Trustworthy Software System No. NRF2018NCR-NSOE003-0001, NRF Investigatorship No. NRF-NRFI06-2020-0001, and the National Natural Science Foundation of China under contracts Nos. 61871258 and U1703261 and the National Key Research and Development Project under contracts No. 2016YFB0800403. We gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC) to our research.

REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*.

[2] G. Balakrishnan, F. Durand, and J. Guttag. 2013. Detecting Pulse from Head Motions in Video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 3430–3437.

[3] Belhassen Bayar and Matthew Stamm. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. 5–10.

[4] Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: Driving Visual Speech with Audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., USA, 353–360. https://doi.org/10.1145/258734.258880

[5] P. Buchana, I. Cazan, M. Diaz-Granados, F. Juefei-Xu, and M.Savvides. 2016. Simultaneous Forgery Identification and Localization in Paintings Using Advanced Correlation Filters. In *ICIP*.

[6] Giovanni Cennini, Jeremie Arguel, Kaan Akşit, and Arno Leest. 2010. Heart rate monitoring via remote photoplethysmography with motion artifacts reduction. *Optics express* 18 (03 2010), 4867–75. https://doi.org/10.1364/OE.18.004867

[7] Yupeng Cheng, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Shang-Wei Lin, Weisi Lin, Wei Feng, and Yang Liu. 2020. Pasadena: Perceptually Aware and Stealthy Adversarial Denoise Attack. *arXiv preprint* (2020).

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2017. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. arXiv:cs.CV/1711.09020

[9] Francois Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. 1800–1807.

[10] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 2017. Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection. 159–164.

[11] Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video Face Replacement. *ACM Trans. Graph.* 30, 6 (Dec. 2011), 1–10. https://doi.org/10.1145/2070781.2024164

[12] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. 2019. On the Detection of Digital Face Manipulation. arXiv:cs.CV/1910.01717

[13] Vincent Nozick Darius Afchar. [n.d.]. FaceForensics Benchmark. http://kaldir.vc.in.tum.de/faceforensics_benchmark/.

[14] Vincent Nozick Darius Afchar. [n.d.]. MesoNet. https://github.com/DariusAf/MesoNet/.

[15] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv:cs.CV/1910.08854

[16] Wei Feng, Ruize Han, Qing Guo, Jianke Zhu, and Song Wang. 2019. Dynamic Saliency-Aware Regularization for Correlation Filter-Based Object Tracking. *IEEE TIP* 28, 7 (2019), 3232–3245.

[17] Jessica Fridrich and Jan Kodovsky. 2012. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security* 7 (06 2012), 868–882.

[18] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormählen, Patrick Pérez, and Christian Theobalt. 2014. Automatic Face Reenactment. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 4217–4224.

[19] P. Garrido, Levi Valgaerts, Hamid Sarmadi, I. Steiner, Kiran Varanasi, P. Pérez, and C. Theobalt. 2015. VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track. *Computer Graphics Forum* 34 (05 2015). https://doi.org/10.1111/cgf.12552

[20] Miroslav Goljan and Jessica Fridrich. 2015. CFA-aware features for steganalysis of color images. *Proceedings of SPIE - The International Society for Optical Engineering* 9409 (03 2015).

[21] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez. 2018. Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation, and COTS Evaluation. *IEEE Transactions on Information Forensics and Security* 13, 8 (2018), 2001–2014.

[22] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:stat.ML/1406.2661

[23] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. 2017. Learning Dynamic Siamese Network for Visual Object Tracking. In *ICCV*. 1781–1789.

[24] Qing Guo, Wei Feng, Ce Zhou, Chi-Man Pun, and Bin Wu. 2017. Structure-Regularized Compressive Tracking With Online Data-Driven Sampling. *IEEE TIP* 26, 12 (2017), 5692–5705.

[25] Qing Guo, Ruize Han, Wei Feng, Zhihao Chen, and Liang Wan. 2020. Selective Spatial Regularization by Reinforcement Learned Decision Making for Object Tracking. *IEEE TIP* 29 (2020), 2999–3013.

[26] Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Bing Yu, Wei Feng, and Yang Liu. 2020. Watch out! Motion is Blurring the Vision of Your Deep Neural Networks. *arXiv preprint arXiv:2002.03500* (2020).

[27] Qing Guo, Xiaofei Xie, Felix Juefei-Xu, Lei Ma, Zhongguo Li, Wanli Xue, Wei Feng, and Yang Liu. 2020. SPARK: Spatial-aware Online Incremental Attack Against Visual Tracking. *European Conference on Computer Vision (ECCV)* (2020).

[28] D. Güera and E. J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–6.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. 2019. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing* 28, 11 (2019), 5464–5478.

[31] Javier Hernandez-Ortega, Julian Fierrez, Aythami Morales, and Pedro Tome. 2018. Time Analysis of Pulse-Based Face Anti-Spoofing in Visible and NIR. 657–6578.

[32] Guillaume Heusch and Sébastien Marcel. 2018. Pulse-based Features for Face Presentation Attack Detection. 1–8.

[33] Yihao Huang, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. 2020. FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction. *ACM International Conference on Multimedia (ACM MM)* (2020).

[34] Yihao Huang, Felix Juefei-Xu, Run Wang, Qing Guo, Xiaofei Xie, Lei Ma, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. 2020. FakeLocator: Robust Localization of GAN-Based Face Manipulations. *arXiv preprint arXiv:2001.09598* (2020).

[35] Kenneth Humphreys, Tomas Ward, and Charles Markham. 2007. Noncontact simultaneous dual wavelength photoplethysmography: A further step toward noncontact pulse oximetry. *The Review of scientific instruments* 78 (05 2007), 044304. https://doi.org/10.1063/1.2724789

[36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv:cs.NE/1710.10196

[37] Tero Karras, Samuli Laine, and Timo Aila. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv:cs.NE/1812.04948

[38] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019. Analyzing and Improving the Image Quality of StyleGAN. arXiv:cs.CV/1912.04958

[39] X. Li, J. Chen, G. Zhao, and M. Pietikäinen. 2014. Remote Heart Rate Measurement from Face Videos under Realistic Situations. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 4264–4271.

[40] Xiaobai Li, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikainen. 2016. Generalized face anti-spoofing by detecting pulse from face videos. 4244–4249.

[41] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. 2019. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. arXiv:cs.CV/1904.09709

[42] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. 2018. Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. 389–398.

[43] Lei Ma, Felix Juefei-Xu, Jiyuan Sun, Chunyang Chen, Ting Su, Fuyuan Zhang, Minhui Xue, Bo Li, Li Li, Yang Liu, Jianjun Zhao, and Yadong Wang. 2018. DeepGauge: Multi-Granularity Testing Criteria for Deep Learning Systems. In *The 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)*.

[44] Lei Ma, Felix Juefei-Xu, Minhui Xue, Bo Li, Li Li, Yang Liu, and Jianjun Zhao. 2019. DeepCT: Tomographic Combinatorial Testing for Deep Learning Systems. *Proceedings of the IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)* (2019).

[45] Lei Ma, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Felix Juefei-Xu, Chao Xie, Li Li, Yang Liu, Jianjun Zhao, and Yadong Wang. 2018. DeepMutation: Mutation Testing of Deep Learning Systems. In *The 29th IEEE International Symposium on Software Reliability Engineering (ISSRE)*.

[46] Scott McCloskey and Michael Albright. 2018. Detecting GAN-generated Imagery using Color Cues. arXiv:cs.CV/1812.08247

[47] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Use of a Capsule Network to Detect Fake Images and Videos. arXiv:cs.CV/1910.12467

[48] X. Niu, S. Shan, H. Han, and X. Chen. 2020. RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. *IEEE Transactions on Image Processing* 29 (2020), 2409–2423.

[49] Ewa Nowara, Ashutosh Sabharwal, and Ashok Veeraraghavan. 2017. PPGSecure: Biometric Presentation Attack Detection Using Photopletysmograms. 56–62.

[50] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Frédo Durand, William T. Freeman, and Wojciech Matusik. 2018. Learning-based Video Motion Magnification. arXiv:cs.CV/1804.02684

[51] Online. [n.d.]. 81 Facial Landmarks Shape Predictor. https://github.com/codeniko/shape_predictor_81_face_landmarks/.

[52] Online. [n.d.]. Face Recognition Using Pytorch. https://github.com/timesler/facenet-pytorch/.

[53] Xunyu Pan, Xing Zhang, and Siwei Lyu. 2012. Exposing image splicing with inconsistent local noise variances. *2012 IEEE International Conference on Computational Photography, ICCP 2012* (04 2012).

[54] M. Poh, D. J. McDuff, and R. W. Picard. 2011. Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Biomedical Engineering* 58, 1 (2011), 7–11.

[55] Ming-Zher Poh, Daniel McDuff, and Rosalind Picard. 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express* 18 (05 2010), 10762–74. https://doi.org/10.1364/OE.18.010762

[56] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and I. Echizen. 2017. Distinguishing computer graphics from natural images using convolution neural networks. 1–6.

[57] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *International Conference on Computer Vision (ICCV)*.

[58] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. arXiv:cs.CV/1905.00582

[59] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI Conference on Artificial Intelligence* (02 2016).

[60] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis using Neural Textures. arXiv:cs.CV/1904.12356

[61] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time Expression Transfer for Facial Reenactment. *ACM Transactions on Graphics* 34 (10 2015), 1–14. https://doi.org/10.1145/2816795.2818056

[62] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.

[63] Paul Upchurch, Jacob Gardner, Kavita Bala, Robert Pless, Noah Snavely, and Kilian Weinberger. 2016. Deep Feature Interpolation for Image Content Changes. (11 2016).

[64] Run Wang, Felix Juefei-Xu, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Yang Liu. 2020. Amora: Black-box Adversarial Morphing Attack. *ACM International Conference on Multimedia (ACM MM)* (2020).

[65] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. 2020. DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices. *ACM International Conference on Multimedia (ACM MM)* (2020).

[66] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. 2020. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. *International Joint Conference on Artificial Intelligence (IJCAI)* (2020).

[67] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William T. Freeman. 2012. Eulerian Video Magnification for Revealing Subtle Changes in the World. *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)* 31, 4 (2012).

[68] Xiaofei Xie, Lei Ma, Felix Juefei-Xu, Minhui Xue, Hongxu Chen, Yang Liu, Jianjun Zhao, Bo Li, Jianxiong Yin, and Simon See. 2019. DeepHunter: A Coverage-Guided Fuzz Testing Framework for Deep Neural Networks. In *ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*.

[69] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. 2019. Remote Heart Rate Measurement from Highly Compressed Facial Videos: an End-to-end Deep Learning Solution with Video Enhancement. arXiv:eess.IV/1907.11921

[70] Ce Zhou, Qing Guo, Liang Wan, and Wei Feng. 2017. Selective object and context tracking. In *ICASSP*. 1947–1951.

[71] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. 2018. Two-Stream Neural Networks for Tampered Face Detection. arXiv:cs.CV/1803.11276