

# How Do the Hearts of Deep Fakes Beat?

## Deep Fake Source Detection via Interpreting Residuals with Biological Signals

Umur Aybars Ciftci  
Binghamton University  
uciftci@binghamton.edu

İlke Demir  
Intel Corporation  
idemir@purdue.edu

Lijun Yin  
Binghamton University  
lijun@cs.binghamton.edu

### Abstract

*Fake portrait video generation techniques have been posing a new threat to the society with photorealistic deep fakes for political propaganda, celebrity imitation, forged evidences, and other identity related manipulations. Following these generation techniques, some detection approaches have also been proved useful due to their high classification accuracy. Nevertheless, almost no effort was spent to track down the source of deep fakes. We propose an approach not only to separate deep fakes from real videos, but also to discover the specific generative model behind a deep fake. Some pure deep learning based approaches try to classify deep fakes using CNNs where they actually learn the residuals of the generator. We believe that these manipulation artifacts by disentangling them with biological signals. Our key observation yields that the spatiotemporal patterns in biological signals can be conceived as a representative projection of residuals. To justify this observation, we extract PPG cells from real and fake videos and feed these to a state-of-the-art classification network for detecting the generative model per video. Our results indicate that our approach can detect fake videos with 97.29% accuracy, and the source model with 93.39% accuracy.*

### 1. Introduction

Artificial intelligence (AI) approaches to generate synthetic videos [60, 35, 34] have reduced required level of skill for realistic image manipulation [63, 7]. These advancements precipitated the rise of deep fakes [5, 9], synthetic portrait videos of real humans, photorealistic enough to be used as fakes. Although this technology has been developed with positive intent for movies [6, 10], advertisement [2], virtual clothing [70], and entertainment [4]; unfortunately this strong impact attracted malicious users to exploit deep fakes for political misinformation [11] and pornography [3].

This threat to information integrity has consequences in privacy, law, politics, security, and policy, and has the potential to form a social erosion of trust [22]. As a defense mechanism, deep fake detection methods have been introduced [67, 48, 40], which define the problem as a binary classification. It is conceivable that, as more realistic and complex generation methods are developed over time, detection methods should also have a more profound development and a deeper understanding. Deciding the authenticity of a video is demanded, however finding the source is even more important and challenging for tracking, prevention, and combating their spread. We propose a deep fake source detector that predicts the source generative model for any given video. To our knowledge our approach is the first to conduct a deeper analysis for source detection that interprets residuals of generative models for deep fake videos.

Biological signals are present in all humans. Anatomical actions such as heart beat, blood flow, or breathing, create subtle changes that are not visible to the eye but still detectable computationally. For example, when blood moves through the veins, it changes the skin reflectance over time, due to the hemoglobin content in the blood. Approaches to extract photoplethysmography (PPG) signals are developed to recognize such changes by image processing techniques.

As of now, no generative model is able to create deep fakes with consistent PPG signals. Several previous approaches utilize similar biological signals to detect 3D synthetic CG renders([24]) and deep fakes ([23, 40, 68]). [23] in particular proves that spatiotemporal inconsistency of biological signals can be exploited to detect deep fakes. Our key observation follows the fact that biological signals are not yet preserved in deep fakes, and those signals produce different signatures in terms of the generative noise. Thus, we can interpret biological signals as a projection of the residuals in a known dimension that we can explore to find the unique signature per model. This motivates us to utilize these signals for the recognition of the current and future generative models behind all deep fake videos. More importantly, the source detection can also improve the overall fake detection accuracy, because real videos with inconsistent bi-

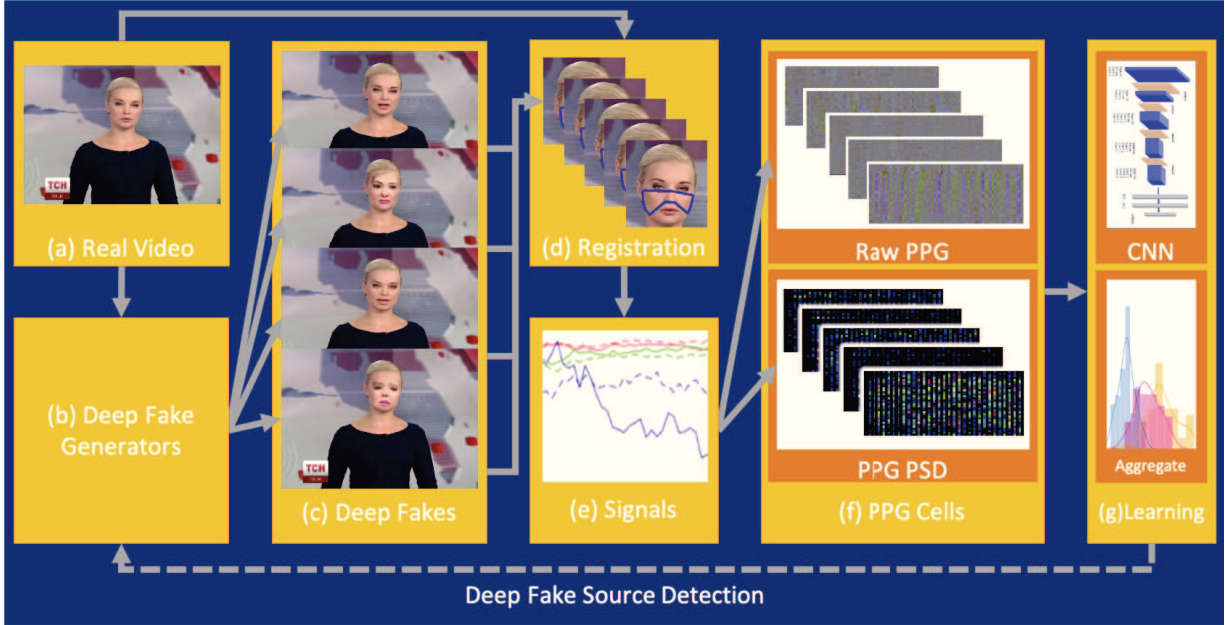


Figure 1. **Overview.** From real videos (a), several generators (b) create deep fakes with residuals specific to each model (c). Our system extracts face ROIs (d) and biological signals (e), to create PPG cells (f) where the residuals are reflected in spatial and frequency domains. Then it classifies both the authenticity and the source of any video (c) by training on PPG cells and aggregating window predictions (g).

ological signals (due to occlusions, illumination changes, etc.) can still be recognized as real videos as they do not conform to the signature of any generative model.

In our work, we extract 32 raw PPG signals from different locations in the face from a window of frames, from a video of windows. We then encode the signals along with their spectral density into a spatiotemporal block, which is so-called *PPG cell*. We feed PPG cells into an off-the-shelf neural network to recognise the signatures of the distinct residuals of the source generative models. Lastly, we combine per sequence predictions into a per video prediction using average log of odds [32]. Our approach achieves the prediction for the authenticity of the video by 97.29%, and the generative model by 93.39% on FaceForensics++ [53] dataset. We evaluate our approach on five datasets with multiple [53], single [42, 68, 52], and unknown [23] generators, against five state of the art source models, and seven backbones. We also conduct an ablation study on various setups for comparison, and compel our approach towards extension to new models and detecting unseen generators.

In summary, the contributions of this paper are listed as,

- a novel approach for deep fake source detection, leading deep fake detection research to a new perspective,
- a new discovery that the projection of generative noise into biological signal space can create unique signatures per model, and
- an advanced general deep fake detector that can outperform current approaches in fake/real classification, while also predicting the source generative model.

## 2. Related Work

### 2.1. Generative Models for Deep Fakes

There exist various deep fake methods in the literature [5, 62, 60, 61, 9, 7, 35, 67]. We categorize these methods broadly based on their face synthesis as (i) face generation, (ii) face reenactment, and (iii) face manipulation techniques. The first category for face generation mostly consists of generative adversarial network (GAN) based methods. For example, StyleGAN [35] and ProGAN [34] are methods to create entire fake faces. The second category for face reenactment includes the face replacement methods using model warping or swapping techniques, for example using a 3D model of another person such as [5, 62, 60, 7, 51, 61, 27, 66]. The third category for face manipulation mostly focuses on facial expression transfer or mouth shape and movement synthesis from lip reading, while keeping the face identity intact [61].

### 2.2. Deep Fake Detectors

For fake image detection from the face generation category, several typical signatures have been identified including saturation cues [46], frequencies of generated images for fingerprints of GAN models [69], and discrete cosine transformation residuals [54].

For the facial reenactment, detection is usually performed per frame, which also utilizes temporal information. To search for some artifacts which may occur due to

the facial differences between the source and target faces, Boulkenafet et al. [18] estimate distortions in the generated faces, Barin et al. [15] investigate compression artifacts, Yang et al. [68] recognize inconsistent head poses, Li et al. [40] detect blinking effects, and Li and Lyu [41] search for face warping artifacts in the generated faces successfully. In addition, other markers such as biological signals [23] and lighting inconsistency [57] have also been explored. Specific generative networks [12, 59, 29, 36, 71, 72, 50] have been applied as the discriminators.

Similarly, above methods can be used for detection of face manipulation tasks. Beyond that, motion and extra modality can also be used as auxiliary components to facilitate detection, e.g., inconsistent mouth movements [45] and [38], and audio/visual verification [37] and [39].

### 2.3. Source Detectors

To our knowledge, most existing deep fake source detectors are image-based, and exploit various attributes of synthetic imagery such as GAN model fingerprints [69, 25, 44, 28], camera patterns [43], or image attribution [13]. Yu et al. [69] identify fully synthetic images that are generated with ProGAN [34], SNGAN [47], CramerGAN [16], MMDGAN [17] and analyze their fingerprints through frequency analysis. Lukas et al. [43] and Cozzolino et al. [25] analyze camera sensor noise for natural images. Marra et al. [44] find GAN residual fingerprints in final synthetic image patterns. Albright and McCloskey [13] examine the source camera attribution on GANs. Although all of those approaches can be regarded as an interpretation of the generative noise in a specific GAN generated image domain, they have not been evaluated on deep fake videos yet, neither have they been applied to any of following domains, such as deep fake videos, or biological signals.

### 2.4. Deep Fake Datasets

With the increase of deep fake generation, the exigencies of detecting such doctored data become essential. While a large amount of such videos/images spread on internet or social media, it is highly demanded to have benchmark datasets specifically curated for research of deep fake detection. With respect to the data source generation, we categorize the existing datasets by two types: (1) datasets with single model generation and (2) datasets using multiple generative sources. Although there exist several synthetic face image datasets [35, 1, 49], here we focus on video datasets due to the absence of biological signals in single images.

The majority of existing deep fake video datasets contain videos created by single, easy-access, and popular generative sources. For instances, UADFV [68] dataset contains 48 real and 48 fake videos generated by FakeAPP [9]. DeepfakeTIMIT [39] dataset has 650 deep fake videos generated using faceswap-GAN [8] where vidtimit [55] videos

are used as originals. FaceForensics [52] dataset congregates 1,004 videos from the internet with their deep fake versions created by Face2Face [62], resulting in 2,008 videos. Celeb-DF [42] dataset collects 590 real videos of famous actors, with 5,639 deep fake versions generated by an improved synthesis process [42].

A typical dataset generated by multiple generative methods is the commonly used FaceForensics++ [53] (FF) dataset, which includes 1,000 real videos and 4,000 fake videos, generated by four generative models – FaceSwap [7], Face2Face [62], Deepfakes [5], and Neural Textures [60]. Recently, an in-the-wild deep fake dataset was created by Ciftci et al. [23], in which 140 videos are collected online, and half of them are fake. However, source models of those fake videos are unknown, thus posing a big challenge for in-the-wild deep fake source detection.

## 3. PPG Cells

Biological signals have been proven as an authenticity indicator for real videos, which have been used as a distinguishable biomarker for deep fake detection [23]. As we know, a synthetic person shown in a fake video does not exhibit a similar pattern of heart beat as the one shown in a real video does [23]. *Our key finding emerges from the fact that we can interpret these biological signals as fake heart beats that contain a signature transformation of the residuals per model.* Thus, it gives rise to a new exploration of these biological signals for not only determining the authenticity of a video, but also classifying its source model that generates the video. Our proposed system for detection of both deep fakes and their sources is outlined in Figure 1.

In order to capture the characteristics of biological signals consistently, we define a novel spatiotemporal block, called the PPG cell. The PPG cells combine several raw PPG signals and their power spectra, extracted from a fixed window. The generation of PPG cells starts with finding the face in every frame using a face detector [14]. In case the window contains multiple faces, we process signals individually and aggregate the results in the final step.

The second step is to extract regions of interests (ROI) from the detected faces that have as much stable PPG signals as possible (Figure 1(d)). Biological signals are sensitive to facial movements, illumination variations, and facial occlusions. In order to extract these areas robustly, we use the face region between eye and mouth regions, maximizing the skin exposure. As the PPG signals from different face regions are correlated with each other [23], locating the ROIs and measuring their correlation become a crucial step to enhance the detection.

The third step involves aligning these nonlinear ROIs to a rectangular image. We employ Delaunay triangulation [26], followed by a nonlinear affine transformation per triangle to transform each triangle into the rectified image.

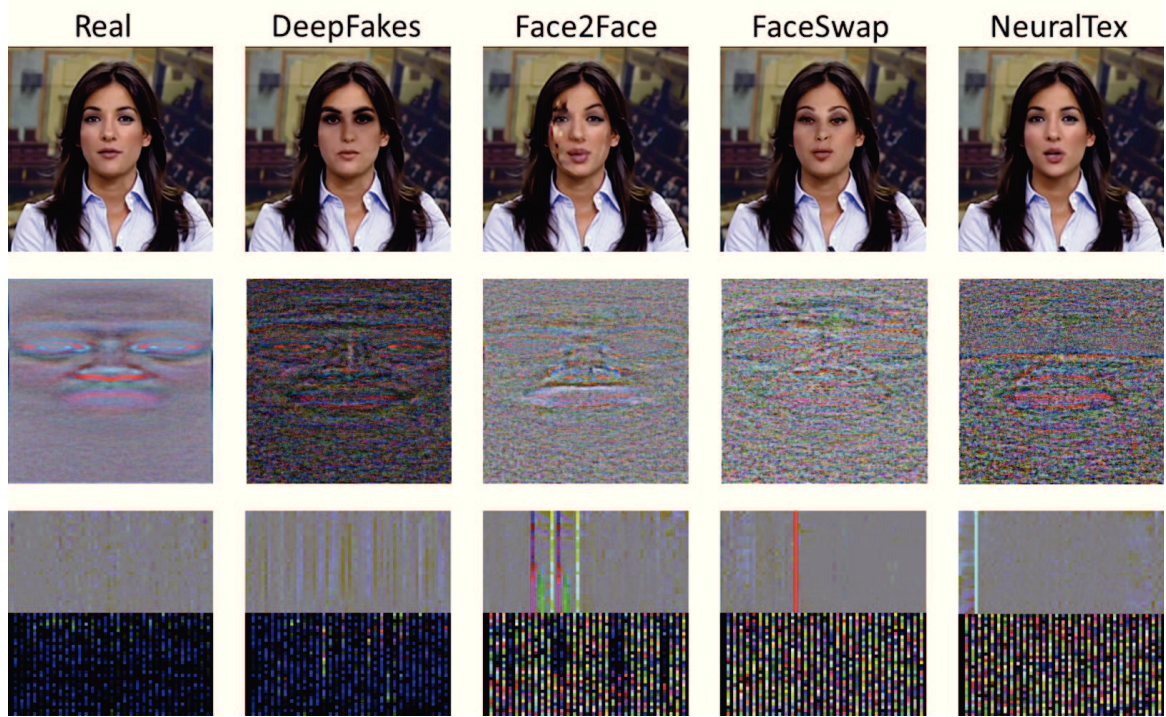


Figure 2. **PPG Cells**. Example frames per  $\omega = 64$  window (top), and their PPG cells (bottom) consisting of raw PPG and PPG PSD, of a real video (left) and its deep fakes per generative model (rest). Middle row represents an approximation to the accumulated residuals over all videos, which correlates with the colors in the PPG spectra.

In the fourth step, we divide each image into 32 equal-size squares and calculate the raw Chrom-PPG signal per square in a fixed window with the size of  $\omega$  frames, without interruptions in face detection (Figure 1(e)). Then, we calculate the Chrom-PPG in the rectified image [65] since it produces more reliable PPG signals [64, 65]. For each window we now have  $\omega$  times 32 raw PPG values. We reorganize these into a matrix of 32 rows and  $\omega$  columns, forming the base of the PPG cells as shown in Figure 1(f) and Figure 2 top half of bottom rows. Note that the bright columns correspond to significant motion or illumination changes where the PPG signal deviates abruptly.

The final step adds the information from the frequency domain to the PPG cells. We calculate the power spectral density of each raw PPG value in the window and scale it to  $\omega$  size. We concatenate the power spectra to the bottom to generate PPG cells with 64 rows and  $\omega$  columns (Figure 1(f)). Figure 2 bottom row shows example PPG cells of deep fakes generated from the same window, with an example frame from each window at the top row. To analyze the contribution of the spectral information, we conduct experiments on PPG cells both with and without this last step and compare their accuracies (Section 6.2).

Having defined PPG cells, now we can demonstrate a claim for our main hypothesis: the projection of residuals

of deep fake generators into the biological signal domain creates a unique pattern that can be utilized for source detection. As proposed by [49], GAN residuals can be approximated by consistent noise in fake images. We apply temporal non-local means denoising on the aligned face in one frame from each video in FF. We then accumulate and normalize the difference of original and denoised images, and subtract the noise of the real images from each corresponding fake residual to obtain the middle row in Figure 2, containing the “fingerprint” per generator. For the real class, we demonstrate the overall noise accumulation. The colors of PPG-PSD correspond to different frequencies in the spectra of these residuals, and some of these frequencies are actually visible in the residual accumulation images. Our main observation follows this correlation between the residuals and our PPG cells: residuals create unique variations in the “deep fake heart beats” per model.

#### 4. Model Architecture

As introduced in the related work section, the state of the art fake detection approaches employ binary classification techniques. For this binary classification task, even shallow CNNs are demonstrated to be useful with the addition of biological signals [23] when compared to complex network architectures without biological signals. However, as we

take one step further from these approaches and introduce multiple classes for source detection, we need a more complex feature space segmentation, thus we put more emphasis on the deep learning model architecture. We formulate this as a multi-label classification task with equally probable classes of different generative sources and real videos.

Our learning setting is built on the FaceForensics++ (FF) dataset with a 70%-vs-30% split, where we generate PPG cells with a window size of  $\omega = 128$ . FF dataset contains 4 different generative models, and we add real videos as the fifth class. Using a simple CNN with 3 VGG [56] blocks, we achieve 68.45% accuracy for PPG cell classification on 5 classes in the FF dataset, showing the need for a higher capacity model. Extending with another VGG block results in 75.49%, confirming our intuition. Both to follow this intuition and also to keep our implementation simple, we experiment with VGG16 [56], VGG19 [56], InceptionV3 [58], Xception [20], ResNet50 [30], DenseNet201 [33], and MobileNet [31], training for 100 epochs, with  $\omega = 128$ , using the same 70%-vs-30% split. Table 1 lists the results of PPG cell classification on the test set, where VGG19 achieves the highest accuracy for differentiating the 4 different generative models and real videos of FF (Figure 1(f)). Complex networks like DenseNet and MobileNet overfit, reaching a very high training accuracy, but failing on the test set.

Backbone	FD Accuracy	SD Accuracy
ResNet50	19.31%	52.23%
MobileNet	27.26%	33.16%
Inception	52.81%	58.60%
DenseNet201	30.82%	37.04%
Xception	70.72%	68.54%
VGG16	71.83%	76.94%
VGG19	<b>76.15%</b>	<b>81.06%</b>

Table 1. **PPG Cell Classification Accuracy.** Overall accuracies with different models for fake detection (FD) as binary classification and source detection (SD) as multi-class classification with  $\omega = 128$  on FF dataset.

## 5. Video Classification

Even though our  $\omega$ -frames PPG cells can act as mini videos, a full video consists of several windows of PPG cells, depending on its length. Therefore we need to aggregate per-cell predictions into per-video predictions. Instead of brute force majority voting, we exploit the prediction confidences and employ log of odds to output the final video accuracies (Figure 1(f)). We document different voting schemes for this process in Table 2, where we compare majority voting, highest average probabilities, two highest average probabilities, and average of log odds on our cell prediction results by VGG19 using  $\omega = 128$ . Average logits increase the video source classification accuracy to 84.93%, 0.46% higher than majority voting, as it is

more robust against outliers by utilizing all predictions for all classes of PPG cells for a given video. We would like to conclude this section by noting that the longer the video is, the more PPG cells we have, and the stronger predictions our system will make, based on this aggregation process.

Aggregation	Video SD Accuracy
majority voting	84.47%
$\langle \rho \rangle$ , where $\rho > 50\%$	83.53%
$\rho_{max}$	83.60%
$\langle \{\rho_{max_1}, \rho_{max_2}\} \rangle$	83.19%
$\langle \log \frac{\rho}{1-\rho} \rangle$	<b>84.93%</b>

Table 2. **Prediction Aggregation from PPG Cell Classification.** Video source detection accuracies based on different voting schemes for the prediction probabilities ( $\rho$ ).  $\langle . \rangle$  denotes the mean.

## 6. Results

Our system is implemented in python utilizing OpenFace [14] library for face detection, OpenCV [19] for image processing, and Keras [21] for neural network implementations. Most of the training and testing is performed on a desktop with a single NVIDIA GTX 1060 GPU, with tractable training times. The most computationally expensive part of the system is the extraction of PPG cells from large datasets, which is a one time process per video. In this section we document our analysis, results, and some ablation studies. Unless otherwise noted, we set our testbed as the FF dataset with the same 70%-vs-30% split – 700 real videos and 4\*700 deep fakes for training, and 300 real videos and 4\*300 deep fakes for testing.

### 6.1. Source Classification Accuracy

To better evaluate our video source classification, we analyze how uniquely each generative model is detected using the biological signals as a modulator for residuals. This analysis supports our claim of different generative models having signature patterns projected to the biological signal space. As per Figure 3, our approach correctly detects real videos with 97.3%, and generative models with at least 81.9% accuracy for five classes (1 real and 4 fakes) of FF.

### 6.2. Ablation Study

In this section, first we train and test on different setups, namely (i) without real videos in the training set, (ii) without the power spectrum in PPG cells, (iii) without biological signals, and (iv) using full frames instead of face ROIs, where  $\omega = 64$  and FF dataset split are set as constants. In the second part, we analyze the effect of  $\omega$ .

#### 6.2.1 Different Setups

Comparing the first two columns in Table 3, overall accuracy very slightly increases, which may be expected when



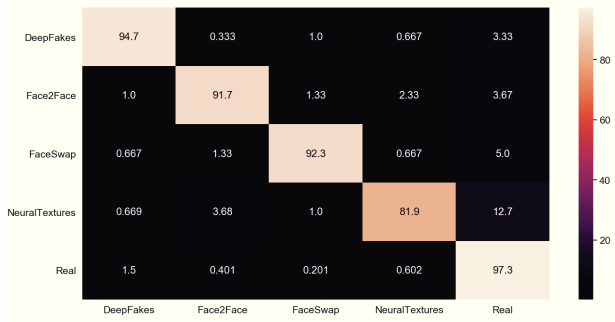


Figure 3. **Confusion Matrix for Class Accuracies.** Video source detection accuracies per 4 generative models and real videos with  $\omega = 64$  on the FF test set, with an average of 93.39% accuracy.

there are less number of classes. Since there exist several binary deep fake detectors, this test ensures that our method can be used as a secondary step to detect the source, after a video is determined to be fake.

Source	All	-real	-PSD	-PPG	Full
DeepFakes	94.7	94.66	94.66	93.00	57.33
Face2Face	91.7	91.66	94.66	87.33	37.33
FaceSwap	92.3	94.00	94.66	92.66	45.66
NeuralTex	81.9	93.07	85.95	83.61	41.33
Real	97.3	NA	89.66	87.20	51.00
Total	93.39	93.57	92.11	88.77	46.53

Table 3. **Ablation Study.** Video source detection accuracies without reals, without PSD part of PPG cells, without biological signals, and on full frames (not only faces).

Comparing the first column to the third column, overall increase is only 1.28% in accuracy. However detection of real videos has an increase of 7.64%, which confirms the main contribution of the power spectrum: the spatiotemporal correlation of biological signals in real videos is not preserved in deep fakes, so it is useful in authenticity detection. The last two columns re-justify the contributions of [23] that (i) biological signals are a crucial factor in fake detection, and (ii) training on faces instead of full frames improves the accuracy.

### 6.2.2 Window Length

The duration from which to extract PPG signals plays an important role in the stability and representative power of the PPG cells. Short windows may miss PPG frequencies and long windows may include too much noise to overshadow the actual signal. We test our method with different window sizes of  $\omega = \{64, 128, 256, 512\}$  frames to balance these ends, on the same setup discussed before (Table 4). With an optimum of  $\omega = 64$ , as we increase the window length, PSNR decrease, and the accuracy drops.

Source	$\omega = 64$	$\omega = 128$	$\omega = 256$	$\omega = 512$
DeepFakes	94.66	93.62	93.26	88.99
Face2Face	91.66	87.62	85.23	69.29
FaceSwap	92.33	90.96	83.94	78.26
NeuralTex	81.93	69.23	84.56	31.11
Real	97.29	83.27	82.88	78.89
Total	<b>93.39</b>	84.93	85.97	73.75

Table 4. **Effect of Window Length.** Video source detection with varying  $\omega$  frame windows.

### 6.3. Extending with New Models

Although we have detected four generative models, it is still a challenging task as new deep fake sources emerge rapidly. To justify that our approach can extend to new models, we combine our FF setup with the single generator dataset CelebDF[42] and repeat the analysis. We randomly select 1,000 fake videos from CelebDF and create a sixth class for their generative model. Our approach achieves 93.69% overall accuracy with 92.17% accuracy on CelebDF, concluding that we can adapt to new models (Table 5). We emphasize that we do not need real counterparts, we only train on fake samples.

Source	Video SD Accuracy
CelebDF	92.17%
DeepFakes	94.66%
Face2Face	91.66%
FaceSwap	92.66%
NeuralTex	86.62%
Real	96.89%
Total	93.69%

Table 5. **New Model Extension.** Video source detection accuracies with 1000 fakes from CelebDF added to FF, as a new class.

This experiment also amplifies our motivation of detecting generative models from their residuals only. In contrast to other source detection methods utilizing the generator architecture or last layers for residual classification, we easily extend to new models without the need for the model specification or the real counterparts of the fake samples.

### 6.4. Comparison

To our knowledge, this paper leads the deep fake detection research towards source detection, utilizing biological signals to classify the residuals of generative models. Some image-based fake detection approaches have been proposed, however, there is no previous work that classifies deep fakes videos using biological signals that enables us to perform a one-to-one comparison with. Thus, we perform experiments with existing approaches on the aforementioned face ROIs, such as (i) the same architecture on video frames (without biological signals), (ii) the same architecture on face ROIs, and (iii) frame-based classification

approaches on face ROIs.

Table 6 lists the accuracies of different models on the test set. The first row uses the same backbone (VGG19), but only on frames, without biological signals. In order to keep the training time tractable, we utilize every 20<sup>th</sup> frame. This can be thought as a baseline for frame-based detection. The second and third blocks are trained and tested on the same dataset, but on segmented and aligned face images. As generators are only swapping or modifying the faces, this approach both makes the training more efficient and improves the accuracies significantly. In this case, our method outperforms even the most complex network, Xception [20], with more than 10% accuracy.

Models	Video SD Accuracy
VGG19 (frames)	46.53%
VGG19 (faces)	76.67%
ResNet50	63.25%
ResNet152	68.92%
Inception	79.37%
DenseNet201	81.65%
Xception	83.50%
Ours	<b>93.69%</b>

Table 6. **Comparison.** Video source detection accuracies on FF dataset, of several models with frame and face based training.

It is worth noting that our approach is advantageous in computational efficiency. As compared to the frame-based detection approach with the same architecture, which takes 29 hours 24 minutes and 43 seconds to train only one epoch with 1.8 million frames in FF on a single GPU, our approach takes only 2 hours and 35 minutes for training the system for 100 epochs. Such a computational efficiency in training and testing with large datasets makes our approach much more feasible to many application fields without demanding high-end computation powers.

## 6.5. Unseen Generators

Previously, we discussed that removal of real class improves the accuracy of finding the distinct residuals of the generative models. This emerges from the fact that PPG signals are affected not only by the generative model residual, but also environmental effects such as lighting, facial movement, and occlusion. As such random artifacts cannot create a pattern, all of those PPG deviations are classified as real, as real is the “chaotic” state without an exact signature. In order to test this hypothesis, we congregate a new test dataset from UADFV [68], FaceForensics [52], Deep Fakes Dataset [23], and CelebDB [42] where we gather 48 real and fake video pairs from each (equal to the smallest of these datasets). On this new collection of 384 videos, we run our previous best model trained on FF with  $\omega = 64$ , with and without the real class.

In addition to the confusion matrix in Figure 3, we depict these new classifications in Figure 4. To begin with, true positives for reals are 100%, 93.61%, 97.82%, and 95.83% according to column 5, rows 1, 3, 5, and 7 respectively. Confirming our hypothesis, UADFV and CelebDF classifications are expected to tend towards the real class (col 5, rows 2&4), because the model does not recognize their signature yet. FaceForensics is expected to be classified as Face2Face [62] (col 2, row 6), as its generative model is within FF. Deep Fake Datasets should have a variety of classification results (row 4) as it contains in-the-wild videos with unknown generators.

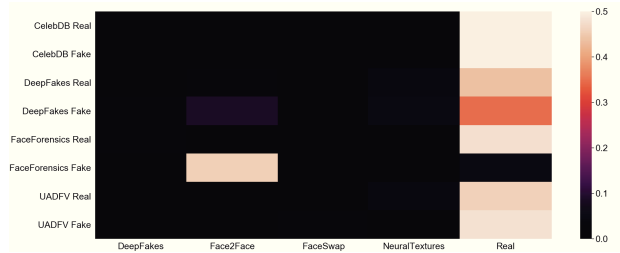


Figure 4. **Unseen Datasets.** Fake and real classification of 96\*4 videos from single and multi source unseen datasets.

## 7. Conclusion

In this paper, we present a deep fake source detection technique via interpreting residuals with biological signals. To our knowledge this is the first method to apply biological signals to the task of deep fake source detection. In addition we experimentally validate our method through various ablation studies. In our experiments we achieve 93.39% accuracy on FaceForensics++ [53] dataset on source detection from four deep fake generators and real videos. Moreover, we demonstrate the adaptability of our approach to new generative models, keeping the accuracy unchanged.

Following the study in biological signal analysis on deep fake videos, the ground truth PPG data along with real and fake videos can enable a novel direction in research on deep fake analysis and detection. In the next stage of the work, we plan to create a new dataset with ground truth PPG, with certain source variations as well as distribution variations.

It is worth noting that our work looks for generator signatures in deep fakes, while the existing work reported by Ciftci et al. [23] looks for signatures in real videos. Theoretically, a holistic system combining these two perspectives can be developed with a jointly trained model for detecting signatures on both authentic and fake videos. We pose this idea as our immediate future work.

## Acknowledgment

This work is supported in part by the National Science Foundation under grant CNS-1629898.

## References

- [1] 100,000 faces generated by ai, 2018. <https://generated.photos>. Accessed: 2020-05-27.
- [2] Are deepfakes the future of advertising? <https://gritdaily.com/deepfakes-in-advertising/>. Accessed: 2020-05-27.
- [3] Deepfake porn nearly ruined my life. <https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/>. Accessed: 2020-05-27.
- [4] Deepfake technology in the entertainment industry: Potential, limitations and protections. <https://amt-lab.org/blog/2020/3/deepfake-technology-in-the-entertainment-industry-potential-limitations-and-protections>. Accessed: 2020-05-27.
- [5] Deepfakes. <https://github.com/deepfakes/faceswap>. Accessed: 2020-03-16.
- [6] Deepfakes are being used to dub adverts into different languages. <https://www.newscientist.com/article/2220628-deepfakes-are-being-used-to-dub-adverts-into-different-languages/>. Accessed: 2020-05-27.
- [7] Faceswap. <https://github.com/MarekKowalski/FaceSwap>. Accessed: 2020-03-16.
- [8] Faceswap-gan. <https://github.com/shaoanlu/faceswap-GAN>. Accessed: 2020-03-16.
- [9] Fakeapp. <https://www.malavida.com/en/soft/fakeapp/>. Accessed: 2020-03-16.
- [10] Here's harrison ford starring in 'solo' thanks to deepfakes. <https://www.popularmechanics.com/culture/movies/a23867069/harrison-ford-han-solo-deepfakes/>. Accessed: 2020-05-27.
- [11] Lawmakers warn of 'deepfake' videos ahead of 2020 election. <https://www.cnn.com/2019/01/28/tech/deepfake-lawmakers/index.html>. Accessed: 2020-05-27.
- [12] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, Dec 2018.
- [13] M. Albright and S. McCloskey. Source generator attribution via inversion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [14] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.
- [15] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro. Aligned and non-aligned double jpeg detection using convolutional neural networks. *J. Vis. Comun. Image Represent.*, 49(C):153–163, Nov. 2017.
- [16] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The cramer distance as a solution to biased wasserstein gradients. *CoRR*, abs/1705.10743, 2017.
- [17] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.
- [18] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, Aug 2016.
- [19] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [20] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [21] F. Chollet et al. Keras. <https://keras.io>, 2015.
- [22] D. Chu, I. Demir, K. Eichensehr, J. G. Foster, M. L. Green, K. Lerman, F. Menczer, C. O'Connor, E. Parson, L. Ruthotto, et al. White paper: Deep fakery – an action plan. Technical Report <http://www.ipam.ucla.edu/wp-content/uploads/2020/01/Whitepaper-Deep-Fakery.pdf>, Institute for Pure and Applied Mathematics (IPAM), University of California, Los Angeles, Los Angeles, CA, Jan. 2020.
- [23] U. A. Ciftci, I. Demir, and L. Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2020.
- [24] V. Conotter, E. Bodnari, G. Boato, and H. Farid. Physiologically-based detection of computer generated faces in video. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 248–252, 2014.
- [25] D. Cozzolino and L. Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020.
- [26] S. Fortune. Handbook of discrete and computational geometry. chapter Voronoi Diagrams and Delaunay Triangulations, pages 377–388. CRC Press, Inc., Boca Raton, FL, USA, 1997.
- [27] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormahlen, P. Perez, and C. Theobalt. Automatic face reenactment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [28] L. Guarnera, O. Giudice, and S. Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [29] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Nov 2018.
- [30] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.
- [32] C. Hsiao. *Logit and Probit Models*, pages 410–428. Springer Netherlands, Dordrecht, 1996.
- [33] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.
- [34] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.



- [35] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [36] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6, Sep. 2018.
- [37] P. Korshunov, M. Halstead, D. Castan, M. Graciarena, M. McLaren, B. Burns, A. Lawson, and S. Marcel. Tampered speaker inconsistency detection with phonetically aware audio-visual features. In *ICML workshop "Synthetic Realities: Deep Learning for Detecting AudioVisual Fakes"*, 2019.
- [38] P. Korshunov and S. Marcel. Speaker inconsistency detection in tampered video. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2375–2379, Sep. 2018.
- [39] N. Le and J.-M. Odobez. Learning multimodal temporal representation for dubbing detection in broadcast media. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 202–206, New York, NY, USA, 2016. Association for Computing Machinery.
- [40] Y. Li, M.-C. Chang, and S. Lyu. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv e-prints*, page arXiv:1806.02877, Jun 2018.
- [41] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [42] Y. Li, P. Sun, H. Qi, and S. Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, United States, 2020.
- [43] J. Lukas, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006.
- [44] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511, 2019.
- [45] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, Jan 2019.
- [46] S. McCloskey and M. Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4584–4588, 2019.
- [47] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *CoRR*, abs/1802.05957, 2018.
- [48] M. S. Nadeem, V. N. L. Franqueira, X. Zhai, and F. Kurugollu. A survey of deep learning solutions for multimedia visual content analysis. *IEEE Access*, 7:84003–84019, 2019.
- [49] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. P. Proença, and J. Fierrez. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 2020.
- [50] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019.
- [51] Y. Nirkin, I. Masi, A. Tran Tuan, T. Hassner, and G. Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 98–105, 2018.
- [52] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. *arXiv e-prints*, page arXiv:1803.09179, Mar 2018.
- [53] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. Faceforensics++: Learning to detect manipulated facial images. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [54] A. Roy, D. Bhalang Tariang, R. Subhra Chakraborty, and R. Naskar. Discrete cosine transform residual feature based filtering forgery and splicing detection in jpeg images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [55] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In M. Tistarelli and M. S. Nixon, editors, *Advances in Biometrics*, pages 199–208, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
- [56] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [57] J. Straub. Using subject face brightness assessment to detect ‘deep fakes’ (Conference Presentation). In N. Kehtarnavaz and M. F. Carlsohn, editors, *Real-Time Image Processing and Deep Learning 2019*, volume 10996. International Society for Optics and Photonics, SPIE, 2019.
- [58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [59] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, MPS '18, pages 81–87, New York, NY, USA, 2018. ACM.
- [60] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), July 2019.
- [61] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6), Oct. 2015.
- [62] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016.
- [63] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros. Detecting photoshopped faces by scripting photoshop. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

- [64] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017.
- [65] W. Wang, S. Stuijk, and G. de Haan. Living-skin classification via remote-ppg. *IEEE Transactions on Biomedical Engineering*, 64(12):2781–2792, 2017.
- [66] W. Wu, Y. Zhang, C. Li, C. Qian, and C. Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [67] D. Yadav and S. Salmani. Deepfake: A survey on facial forgery technique using generative adversarial network. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 852–857, 2019.
- [68] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019.
- [69] N. Yu, L. S. Davis, and M. Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [70] M. Yuan, I. R. Khan, F. Farbiz, S. Yao, A. Niswar, and M. Foo. A mixed reality virtual clothes try-on system. *IEEE Transactions on Multimedia*, 15(8):1958–1968, 2013.
- [71] Y. Zhang, L. Zheng, and V. L. L. Thing. Automated face swapping and its detection. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, pages 15–19, Aug 2017.
- [72] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839, July 2017.