How Do Deepfakes Move?
Motion Magnification for Deepfake Source Detection

İlke Demir
Intel Labs
Santa Clara, CA
ilke.demir@intel.com

Umur Aybars Çiftçi
Binghamton University
Binghamton, NY
uciftci@binghamton.edu

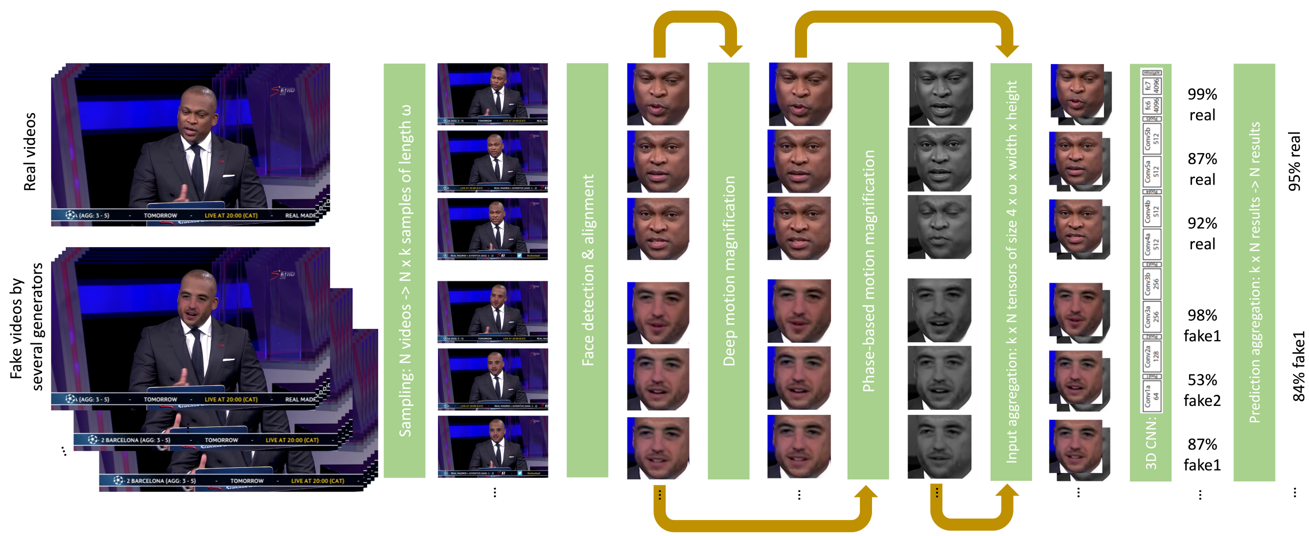

Figure 1. **System Overview.** Starting with real and fake videos from several generators, our approach selects fixed-window samples, extracts and aligns faces, applies deep and phase-based motion magnification to aligned faces, combines magnified outputs, trains a 3D CNN, and aggregates predictions into video predictions to classify if a video is real or its generator.

Abstract

With the proliferation of deep generative models, deepfakes are improving in quality and quantity everyday. However, there are subtle authenticity signals in pristine videos, not replicated by current generative models. We contrast the movement in deepfakes and authentic videos by motion magnification towards building a generalized deepfake source detector. The sub-muscular motion in faces has different interpretations per different generative models, which is reflected in their generative residue. Our approach exploits the difference between real motion and the amplified generative artifacts, by combining deep and traditional motion magnification, to detect whether a video is fake and its source generator if so. Evaluating our approach on two multi-source datasets, we obtain 97.77% and 94.03% for video source detection. Our approach performs at least 4.08% better than the prior deepfake source detector and other complex architectures. We also analyze magnification amount, phase extraction window, backbone network, sample counts, and sample lengths. Finally, we report our results on skin tones and genders to assess the model bias.

1. Introduction

Since the introduction of Generative Adversarial Networks [35] in 2014, deep generative models have been invading the domain of face generation with increasingly photorealistic results. With the advances in transformer and attention-based modules, the control over and the interpretability of such generators are also escalating. The recent Zelensky video [3] spreading misinformation about the Russian invasion, or the debate about Bruce Willis' deepfake rights [1] are just the tip of the iceberg for a desolate digital future where we cannot trust anything we see online [24]. On the other hand, deepfake detection initiatives finally start to take action towards unifying the efforts [5,9].

Deepfake detection research has been historically investigated from two main perspectives: Blind detectors [11, 23, 27, 57, 75] that try to learn the artifacts of fakery by training on several datasets, and prior-based detectors [12, 20, 30, 40, 55, 85] where the authenticity is somehow represented by hidden signals in pristine videos. Blind detectors have the disadvantages of (1) overfitting to the datasets they are trained on and (2) being prone to adversarial attacks [19, 72]. Thus, our approach follows the second perspective towards more generalizable deepfake detectors, where we define the hidden watermark of being human as sub-muscular motion in this paper.

Moreover, prior approaches in deepfake domain solve a simpler task of "is this video real or fake?". Our approach performs **source generator detection**, which is classifying videos into real or several generative model classes used for creating the video. Source detection has been a much less investigated problem than deepfake detection as it goes beyond binary classification. We anticipate that the aforementioned motion cues are representative enough to provide not only the video authenticity, but also the generative model behind a fake video. Although from a research perspective it would make sense to pose this problem only as "which generator created this video?", that question requires prior knowledge about the video being fake. Posing it as "is it fake, and if so, which generator created it?" defines a more relevant and practical source detector in a general setting, enabling any video to be processed without assumptions.

To reveal the real motion and its projection in generative spaces of different models, we use motion magnification. In pristine videos, magnified motion follows the regular human motion with an emphasis, so action units and other muscles are still correlated temporally and spatially. In fake videos, we observe that the generative noise overpowers the sub-muscular motion. Thus, when motion is magnified, generative noise gets amplified instead of the regular human motion patterns. Our approach

- combines traditional and deep motion representations to analyze motion patterns in real and fake videos from different generative sources,
- proposes a novel, robust, and generalizable deepfake source detector based on motion cues, and
- improves both source detection and fake detection, evaluated on two datasets.

Following the motion magnification literature, we combine traditional phase-based magnification [82] which captures small temporal motions and deep magnification [66] which is more robust towards mixed motion patterns. In addition to this dual representation, we employ a 3D CNN variation to train a robust source detector which learns human motion (and its extents) in real videos and amplified generative noise in deepfakes from different source generators. Overview of our approach is depicted in Fig. 1.

We evaluate our deepfake source detector on FaceForensics++ [71] and FakeAVCeleb [47] datasets, obtaining 97.77% and 94.03% source detection accuracies, among 6 and 4 classes, respectively. We compare our source detection results against both complex blind detectors and prior-based detectors, overperforming the best one by 4.08%. To understand the importance of motion magnification components, we conduct several experiments with different magnification levels, simple to complex backbones, different phase-windows, varying sample counts, and for all skin tones. Finally we discuss how it can be deployed in current deepfake detection workflows.

2. Related Work

Deepfake Generation. Deepfakes have been increasing in quality and quantity since the introduction of Generative Adversarial Networks (GANs) [35]. These approaches can (1) generate novel faces from learned distributions [22, 29, 45, 46] mostly in image domain, (2) transfer or modify facial expressions, speech, identity, or mouth movements from a reference motion onto the target faces [68, 78], and (3) swap entire faces from source to target media [4, 7, 53, 77]. Our approach can classify videos created with any of these deepfake generation techniques and our test datasets indeed include generators from each category [4, 6, 7, 43, 53, 68, 77, 79]. To put these generators in context; [6, 79] are graphics based approaches using blendshapes for face transfer, [77] utilizes deep neural textures, [7, 53] are GAN models for face swapping, [68] is a GAN based lip-sync model, [4] is an autoencoder for face swapping, and [43] uses separately trained encoder, synthesis, and vocoder networks for audio generation.

Deepfake Detection. As deepfakes' malevolence starts to impact the society [2, 8, 24], the arms race between generation and detection intensifies [61, 80]. Initial deepfake detection research focus on finding pixel-level artifacts directly from data, proposing "blind" detectors [11, 14, 17, 36, 37, 48, 54, 64, 76, 89, 90, 90]. These approaches tend to learn specific artifacts of the datasets they are trained on, preventing their generalization and domain-transfer to any unseen video. In addition, they are more prone to be affected by adversarial attacks [19, 72].

In contrast, novel deepfake detectors aim to extract unique authenticity signals from real videos as *watermarks of humans*, such as headpose [85], blinks [55], heartbeats [20], eye and gaze properties [30], lighting [74], breathing [50], and other natural, physical, or biological characteristics. While motion-based deepfake detectors emerged recently [34, 60], neither of them can do source detection, uses a dual motion representation, performs cross dataset validation, and [34] is only tested on a small dataset.

The consistency and correlation of these interpretable signals are broken for fake videos, so these approaches provide better generalizability as long as the GAN does not exploit the specific prior as a loss.

Source Detection. As previously defined, source detection tackles the task of identifying the generative model that outputs a synthetic data, only by inspecting the sample. The hidden artifacts that enable source detection, called the generative residue of GAN fingerprints, have first been identified in the patterns of CNN generated images [83]. Since then, several approaches investigate these artifacts in synthetic images, with frequency analysis on 4 GANs [86], in image patterns [59], using latent representations [31], to infer model hyperparameters [15], for camera attributions [13], by sensor noise [58], or to poison GANs [87]. Unfortunately, previous work in this domain investigates images that are fully synthetic, which is not aligned with real world scenarios. Furthermore, most of them assumes that the entire image is AI-generated, in contrast to more traditional deepfakes where only the portion of the image is swapped, sync'ed, or manipulated.

Relatively less work is proposed for videos and only one work proposes source detection on deepfakes [21]. The authors classify deepfakes by their source generator, projecting their generative residue into a biological signal domain. Our approach tackles the same problem of deepfake source detection, however we propose that motion artifacts are more representative (for pristine videos) and more fragile (for fake videos) in the context of generative fingerprints.

Deepfake Datasets. Several video datasets have been proposed for deepfake detection research, we categorize these as single-, multi-, and unknown-source datasets. Image datasets are skipped as there is no motion in single images. Single-source deepfake datasets are created by easy-access GANs and include UADFV [85], Deepfake-TIMIT [52], FaceForensics [70], Celeb-DF [56], and DeeperForensics [44]. These datasets are crucial for deepfake detection, but not for source detection. Multi-source datasets are FaceForensics++ [71] with 5 generators and 6K videos, DFDC [32] with several unknown and undocumented generators and over 100K videos, and FakeAVCeleb [47] with 3 generators and 20K videos. Considering the diversity, consistency, and labeling of the datasets; we select FaceForensics++ (FF) and FakeAVCeleb (FAVC) datasets for training, testing, and evaluation of our approach. Finally, unknown-source deepfake datasets (i.e., in-the-wild deepfakes) have also been proposed [20, 69], which are important for evaluating and understanding model capabilities in an in-the-wild setting. We use the in-the-wild dataset of [20] for cross-model evaluation of our deepfake detector. This validation both acts as a cross-model experiment and as a supporting generalization claim towards unknown methods.

3. Understanding Motion in Deepfakes

Motivated by finding authentic signals in real videos, we follow the discussion of [20] about biological signals. Photoplethysmography (PPG) and Ballistocardiography (BCD) signals are proposed for understanding heart beats of deepfakes, discussing that BCD extraction would require still faces, else the motion of veins would be overpowered by the actual movement. Inspired by this claim, we would like to understand the motion consistency in deepfakes.

Motion magnification is a mature research area with numerous application-specific solutions [26, 51, 65, 84], re-

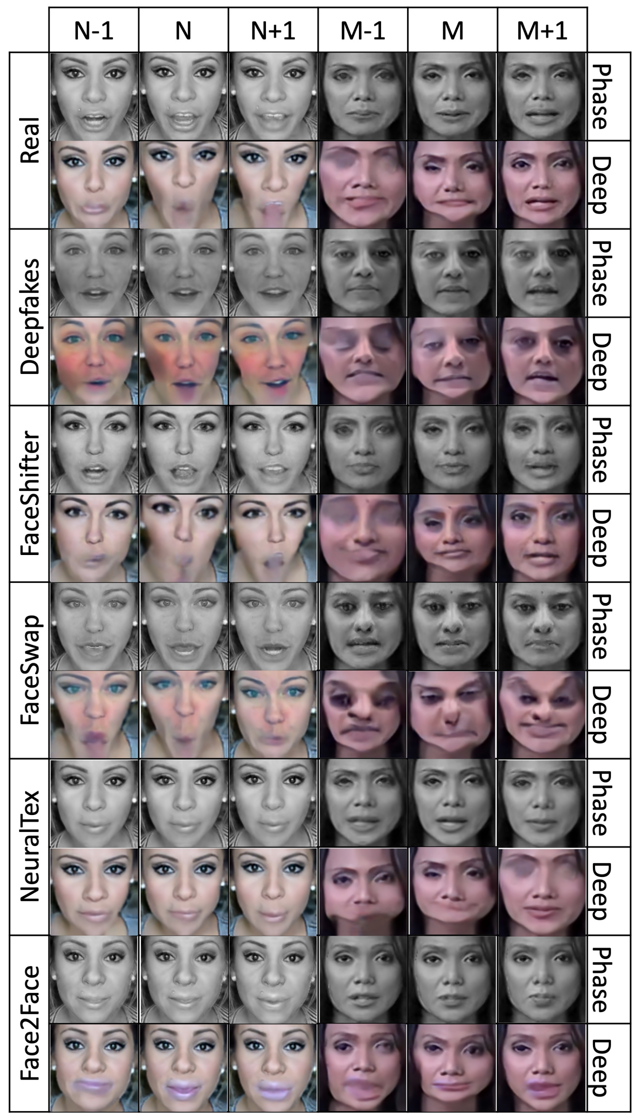

Figure 2. **Motion in Deepfakes.** Each row contains 3 consecutive frames from 2 videos, where motion is magnified by traditional (even rows) or deep (odd rows) methods. Real magnified frames (top 2) are followed by magnified fake frames from 5 generators.

cently extending to deep-learning-based counterparts [66]. Motion magnification has also been explored for deepfake detection recently, obtaining negative results with Euler video magnification [28], without explicit motion magnification [60], and using a two stage CNN+LSTM approach [33]. Unlike prior work focusing on deepfake detection, we claim that, motion discrepancy is useful not only for deepfake detection, but also for source detection, which is a different and harder problem as the next step in the battle against deepfakes. We also claim that, although deep motion magnification learns and models motion robustly, it may not accurately capture smaller motions requiring temporal filters as mentioned in [66], thus, phase-based magnification is also needed for the submuscular motion to be differentiated in real videos. The dual-motion representation strengthens our approach both theoretically and practically (as in Sec. 5.4).

To analyze deepfake motions, we first apply traditional and deep motion magnification to real and fake pairs of videos. Fig. 2 depicts the magnified motion, which is reflected as blurs in deep-motion-magnified frames, are more structural and local in real videos, whereas fake videos experience significant deformations. For the phase-based magnification, we note that the motion is reflected as an accumulation, rather than a blur. This visual observation can also be backed up by comparing the PSNR and SSIM of each real and fake motion-magnified frame. Moreover, different generators (named in the header column) experience this motion dissimilarly as seen in rows 4-13, which supports our main hypothesis of *"motion magnification on deepfakes reveal their source generative model, because the generative noise is amplified as opposed to real motion."*.

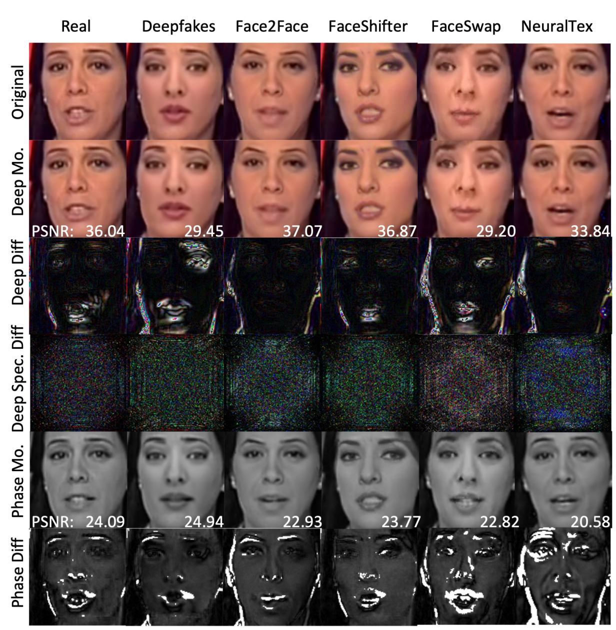

Figure 3. **Quantifying Motion.** Real and 5 corresponding fake images, their deep (row 2) and phase-based (row 5) motion magnifications, differences (rows 3, 6), PSNR, and spectra (row 4).

In Fig. 3, image, frequency, and noise differences of deep and phase-based motion magnified versions of a real image and its five fake variations are shown. We observe that the structure, noise, and distribution of motion differ consistently in all. Real videos have submuscular motion on the cheek, Deepfake videos show asymmetric magnification, and videos that are generated with Neural Textures [77] and Face Swap [6] have magnified boundaries. These amplified differences are also visible in the spectra of magnified motion and serve as the base for our source detector.

4. Motion-based Source Detection

As depicted in Fig. 1, our approach consists of frame sampling, face processing, motion magnification, neural network training, and prediction aggregation.

4.1. Frame Selection

To amplify and understand the motion of the generative residue in deepfake videos, we select k sample intervals of ω frames from each video for training. These samples are selected uniformly from every $(100/k)^{th}$ percentile of the video. The intuition behind this sampling is that videos in these datasets have varying lengths and we do not want any video to dominate the training process. After these fixed samples are gathered, we run face detection on every frame and align faces to extract consistent signals. Each aligned face is fit to a $w \times h$ image to unify the representation. Also, the existence of large actor motions in fake videos may overpower the motion of the residuals, so we tend to sample from lower number of frame sets.

4.2. Motion Magnification

As discussed in [66], phase-based motion magnification may still perform better than deep motion magnification where temporal filters are needed to extract small motions. Thus, we combine both traditional and deep motion magnification by applying them to aligned faces of each k samples of ω frames, obtaining $k \times (\omega - (t-1)) \times 1$ size phase-based magnification output (where t is the frame range parameter of phase-based magnification) and $k \times \omega \times 3$ size deep motion magnification output, per video.

4.2.1 Methodology

Traditionally motion magnification works by decomposing the video into frame representations to magnify the motion by hand crafted filters. In deep motion magnification [66], these filters are learned by a CNN network in three parts. First, the encoder acts as a spatial decomposition filter that extracts a shape representation from consecutive frames and separate texture representations from the shape. Second, the manipulator uses this shape representations to magnify the motion by creating a new magnified shape representation using the shape representations from multiple frames. Finally, the decoder reconstructs the new shape representation with the original texture representations as the motion-magnified frames. This constitutes our deep motion magnification output per video. Phase-based magnification [82], on the other hand, uses an Eulerian approach to motion processing, based on complex-valued steerable pyramids, where their phase variations correspond to local motions in spatial subbands of an image. Phase-based magnification computes local phase variations to measure motion without computing optical flow and performs temporal processing to amplify motion in temporal frequency bands, outputting the second part of our motion-magnified representation.

4.2.2 Parameters

Deep motion magnification uses an amplification factor of m and phase-based magnification uses a sliding window of t frames, thus the output is reduced in length. We merge these two kind of motion magnification outputs into a tensor of $w \times h \times (\omega - (t-1)) \times 4$ for corresponding frames per sample, per video, as the input to our network. We left the discussion on ω, t, and m to our ablation studies in Sec. 5.4.

4.3. Network Architecture

Source detection task is formulated as a multi-class classification problem where n deepfake generators in the dataset plus the originals constitute the class categories. Considering the spatio-temporal nature of our data, we attempt to use transformer-like architectures for source detection. We observe that our motion-enriched representation is powerful enough that transformers easily overfit to our data. Thus, we architect a simpler 3D convolutional neural network, similar to c3d [81]. Our 4D tensors are first input to 64 convolutional kernels of size 3x3x3, followed by batch norm, relu, and maxpool layers; then same block is repeated 4 times with 128, 256, 512, and 512 kernels; followed by two fully connected layers of size 4096 with 0.5 dropout. The selection of this architecture is also backed up by our experiments in Sec. 5.4.

Our dual motion representation relaxes the classification network, so we can use simple and efficient architectures, which significantly reduces the training time. With limited compute resources, for carbon-friendly training, and especially for real-time inference on CPU, it is preferable to use simpler architectures. One can also claim that this reduces the inference time under the assumption that *time(simple network + motion extraction)* $<$ *time(complex network)*.

4.4. Prediction Aggregation

After we obtain results per each sample of each video, we combine k class predictions with their confidences per sample into n video predictions. We experiment with different aggregation techniques in Tab. 7. Providing both segment and video accuracies enables our approach to be suitable for both streaming-based and offline applications.

5. Results

Our approach is implemented in Python utilizing OpenCV [18] for image processing, PyTorch [67] for deep learning, OpenFace [16] for face detection and alignment, vit-pytorch [10], and Efficient-3DCNN [49] libraries for flexible neural network implementations. Most of the training and testing is performed on a desktop with an NVIDIA GeForce RTX 3070, where 100 epochs take a few hours to train. Applying motion magnification is the most computationally expensive part of the system, however, it is an offline task done once per dataset (and for each ablation study with varying motion parameters). Unless otherwise noted, we set $w = h = 112$, $\omega = 16$, $k = 4$, $t = 5$, and $m = 2x$. Phase-based motion magnification frequency coefficients are used as-is from the original paper [82] with $BP = 600\ fps$, $LP = 72\ fps$, and $HP = 92\ fps$ filters. FF [71] is set as the main dataset with the same 70/30 split for all evaluations – 700 real and 700 fake videos from each 5 source generators for training, as a total of 4200 videos for training; and 300 real and 300 fake videos from each 5 source generators, as a total of 1800 videos for testing. FAVC [47] is also used for evaluations (500 real, 700 FaceSwap, 3963 FSGAN, 5014 Wav2Lip videos) with the same split percentages for training and testing.

5.1. Evaluation

The confusion matrices in Fig. 4 demonstrate our source detection accuracy per class. On FF dataset, we obtain 97.77% *video* source detection accuracy, 95.92% *sample* source detection accuracy, and 91% *real class* accuracy. On FAVC, we obtain 94.03% video source detection accuracy, 89.67% sample source detection accuracy, and 91.43% real class accuracy. We emphasize that, our per-class accuracies are much higher for fake classes than the real class, because the model learns the amplified motion of the generative residue. In that sense, real class becomes the “chaotic” class where unknown (or less confident) predictions are pushed into the real class. Real class accuracy (91.43% on FAVC) should not be confused with fake detection accuracy (95.12% on FAVC) as it is produced by a different and complex classification.

	FaceSwap	FSGAN	W2L	Real
FaceSwap	97.53	1.23	1.23	0.00
FSGAN	0.00	100.00	0.00	0.00
W2L	5.05	5.05	84.85	5.05
Real	1.43	1.43	5.71	91.43

	Deepfakes	Face2Face	FaceShifter	FaceSwap	NeuralTex	Real
Deepfakes	100.00	0.00	0.00	0.00	0.00	0.00
Face2Face	0.00	99.33	0.00	0.00	0.00	0.67
FaceShifter	0.00	0.00	99.67	0.00	0.00	0.33
FaceSwap	0.00	0.00	0.00	100.00	0.00	0.00
NeuralTex	0.00	1.33	0.00	0.00	93.00	5.67
Real	0.33	3.33	1.33	1.33	2.67	91.00

Figure 4. **Source Detection Results.** Our approach obtains 94.03% and 97.77% overall video source detection accuracy on FAVC (top) and FF (bottom) datasets, respectively.

5.2. Comparison

In addition to the only other deepfake source detector in the literature [21], we compare our results on FF against complex network architectures used for deepfake detection, in order to emphasize the strength of our dual motion magnification representation in Tab. 1. Our approach beats the best source detector by 4.08% and is much simpler than the deeper networks listed, thus, it has significantly less inference time and it is more generalizable, not over-fitting to specific generators, artifacts, or datasets. We note that source detection is relatively an unexplored area and there is no other method suitable for direct comparison, so we compare with tangential methods doing deepfake detection. Comparing to [60] with 93% fake detection accuracy, which does not perform source detection and uses only phase-based motion magnification, we obtain 97.77% source detection accuracy on the same dataset.

Models	Source Det. Acc.
ResNet50 [38]	63.25%
ResNet152 [38]	68.92%
VGG19 [73]	76.67%
Inception [75]	79.37%
DenseNet201 [41]	81.65%
Xception [23]	83.50%
PPG-based [21]	93.69%
Ours	**97.77%**

Table 1. **Comparison on FF.** Source detection accuracies of several models on FF dataset.

5.3. Cross-model Evaluation

Although cross-model experiments make sense for deepfake detection, there does not exist two multi-source deepfake datasets with the same set of generators to perform a cross-dataset evaluation for source detection. Thus, we assess the generalization of our approach on real class accuracy across datasets. We test our 97.77% model on an in-the-wild dataset [20] (with unknown generators; large motion, illumination, and occlusion artifacts), obtaining **92.64% real class accuracy**. Investigating hard failure cases, large actor motion in deepfakes affects accuracy, whereas other factors are not as relevant. We propose this as the first step to explore open set scenarios with unknown generators, as explored in [25], which enables retraining the model for new generators as their outputs emerge.

5.4. Analysis & Experiments

In this section, we analyze the impact of varying motion parameters t and m, training and testing accuracies of different backbone models, and analyze the accuracies across genders and skin tones.

Magnification	Parameter	Source Det. Acc.
Deep	$m = 2x$	**91.54%**
Deep	$m = 3x$	86.86%
Deep	$m = 4x$	83.16%
Deep	$m = 10x$	74.90%
Phase	$t = 3$	79.88%
Phase	$t = 5$	**85.61%**
Phase	$t = 7$	81.26%
Phase	$t = 10$	82.92%
Phase	$t = 16$	64.85%
Both	$m = 2x$ and $t = 5$	**95.92%**

Table 2. **Motion Magnification Parameters.** Different motion settings for traditional and deep components, with varying magnification coefficient (m) and phase-extraction interval (t).

5.4.1 Analyzing Motion Parameters

In motion magnification literature, the amount of magnification is a significant parameter fine-tuned per application. Over-magnification may lead to complete loss of generative signals, as suspected to be the case in [28]. To investigate this claim, we experiment with several magnification coefficients for deep motion magnification and several window sizes for phase-based motion magnification in Tab. 2. Note that these experiments are done without the dual representation to understand the contribution of each parameter individually. Motion vectors created by generative noise are small, thus we conclude that 2x deep magnification and 5 frame windows for phase-based magnification reveal the sweet spot for emphasizing the motion. As observed from these experiments, only traditional or only deep magnification is not enough to capture generative artifacts, which underlines the contribution of our dual representation.

5.4.2 Visualizing Motion Parameters

In addition to this quantitative analysis, we demonstrate the effects of different parameter values in Fig. 5, for a real video and two deepfakes created from it. We can observe that even for the real video, 10x magnification deteriorates the content. On the other hand, 10-frame phase extraction tends to converge to a mean image of the video, which is not useful either for capturing small motions. Based on these observations and the experiments in Tab. 2, we conclude with $m = 2x$ and $t = 5$ values.

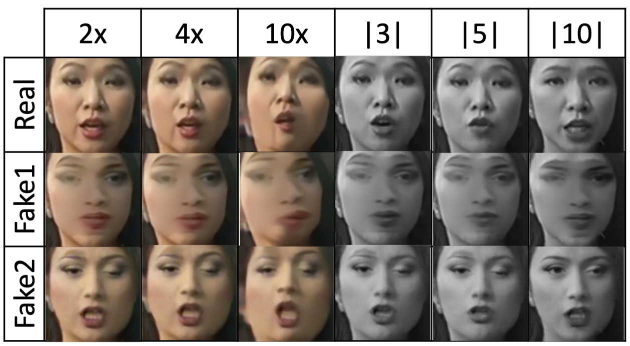

Figure 5. **Magnification Parameters.** Following the experiments on different magnification parameters, we depict the effects of deep motion magnification amount m (left three columns) and phase-based magnification interval t (right three columns).

5.4.3 Backbone Network Analysis

As mentioned in Sec. 4.3, we experiment with different network architectures in accordance with the characteristics of our data 3 and report both training and testing accuracies for source detection. As the motion magnified tensor representation already fortifies the generative artifacts, deeper and more complex networks (like transformers) tend to overfit. In order to observe this phenomenon better, we report the per-sample source detection accuracies before the aggregation step, both for training and testing. We conclude that C3D [81] is powerful enough to robustly learn from the dual-motion representation.

5.4.4 Demographics Analysis

As the last experiment, we want to detect and mitigate any possible racial or gender bias in our dataset or in our algorithm (see [63] for the impact of synthetic data on demographics). To that end, we use the labels in FAVC dataset to report per gender and per skin tone source detection accuracies. We observe that the largest discrepancy in accuracies is between Asian women and American men, with 84.21% and 97.44%. We suspect that this difference may rise from the fact that deepfake generators are not creating such faces with the same fidelity, thus, detection results are also skewed. We also observe that sample detection accuracy is lower for African males, however the aggregation step corrects that. We leave further analysis as future work.

Backbone	Training Acc.	Testing Acc.
Simple3DViT [10]	93.11%	53.56%
3DViT [10]	98.60%	45.97%
CNN-LSTM [62]	95.76%	44.21%
ShuffleNet [88]	98.85%	48.16%
SqueezeNet [42]	99.19%	62.65%
Ours (C3D [81])	99.66%	**95.92%**

Table 3. **Architecture Analysis.** Training and testing accuracies with several architectures for sample source detection to support the strength of our representation and the choice on our backbone.

Skin Tone	Gender	Sample Acc.	Video Acc.
African	Men	79.58%	89.19%
African	Women	96.56%	93.59%
American	Men	95.63%	97.44%
American	Women	89.79%	94.74%
Asian	Men	85.76%	89.74%
Asian	Women	84.25%	84.21%
European	Men	86.46%	92.11%
European	Women	93.12%	94.87%
Indian	Men	90.83%	94.87%
Indian	Women	81.94%	87.18%

Table 4. **Gender & Skin Tone.** We report per sample and per video source detection accuracies on 5 skin tones and 2 genders.

6. Ablation Studies

We experiment with varying number of samples per video (k), changing number of frames in a sample (ω), and different methods for gathering several sample predictions into one video predictions.

6.1. Number of Samples

In order to find optimal parameters, we experiment with changing values for k samples per video. In Tab. 5 we document experiments with $k = \{1, 2, 3, 4\}$, concluding that $k = 4$ is more informative and creates a more diverse dataset, increasing the accuracy. Larger values have incremental contributions within the variance, so we set $k = 4$ with the optimum performance.

k Value	FF Video Acc.
1	95.72%
2	94.83%
3	96.88%
4	**97.77%**

Table 5. **Sample Size Analysis.** k samples per video affects the accuracy. After $k = 4$, contributions are almost constant.

6.2. Sample Length

In order to find optimal parameters, we experiment with ω frame length per sample. In Tab. 6 we document experiments with $\omega = \{4, 8, 16\}$, concluding that $\omega = 16$ is an ideal length where the understanding of temporal motion (per sample accuracy) and the elimination of large motion artifacts (per video aggregation) is balanced. Larger ω values tend to impact the video accuracy by leaking large motion artifacts into the temporal representation.

ω Value	Sample Acc.	Video Acc.
4	91.18%	94.95%
8	91.12%	95.61%
16	**95.95%**	**97.77%**

Table 6. **Sample Length Analysis.** Video samples with ω frames affect the accuracy up to $\omega = 16$.

6.3. Aggregation Methods

We experiment with different aggregation methods to combine k segment predictions into one video prediction in Tab. 7. We choose averaging over other methods since our sample prediction accuracies seem to result higher, as long as there is no large motion or illumination change. Averaging eliminates outliers and grounds the aggregation with respect to the possible artifacts in our videos. Averaging is also a better fit as our sample size is smaller, as opposed to using log of odds, which may work better for longer videos.

Log of Odds	Majority Voting	Averaging
90.61%	97.17%	97.77%

Table 7. **Prediction Aggregation.** Combining sample predictions into video prediction by averaging gives the best accuracy.

7. Conclusion and Future Work

Following several other questions about deepfakes, such as their emotions [39], gazes [30], and hearts [21], we ask *"How do deepfakes move?"*. We propose that motion magnification emphasizes the generative artifacts in deepfakes while preserving pristine motion, which can be used for source detection. Combining deep and phase-based motion magnification, we build a motion-based source detector, achieving accuracies higher than existing source detectors and other complex networks. We support our observations and design choices with ablation studies and experiments, while also performing evaluations on multiple datasets with a cross dataset validation.

In the battle against deepfakes, we believe that source detection plays a crucial role for continuous deployment and integration of detectors into trusted platforms. Emergence of novel generators as well as tracking the malevolent uses of current ones are enabled by source detection, to timely prevent deepfakes causing catastrophic events [3]. Motion as a spatiotemporal signal reflects the sources of these deepfakes and we would like to further analyze and correlate motion with other signals, especially in the multi-modal setting, understanding the relationship of sound, speech, gaze, and gesture with motion.

References

[1] Bruce willis denies selling rights to his face. https://www.bbc.com/news/technology-63106024. Accessed: 2022-11-12. 1

[2] Deepfake porn nearly ruined my life. https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/. Accessed: 2020-05-27. 2

[3] Deepfake video of zelenskyy could be 'tip of the iceberg' in info war, experts warn. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia. Accessed: 2022-11-12. 1, 8

[4] Deepfakes. https://github.com/deepfakes/faceswap. Accessed: 2020-03-16. 2

[5] Deeptrustalliance. https://www.deeptrustalliance.org/. Accessed: 2022-11-12. 1

[6] Faceswap. https://github.com/MarekKowalski/FaceSwap. Accessed: 2020-03-16. 2, 4

[7] Faceswap-gan. https://github.com/shaoanlu/faceswap-GAN. Accessed: 2020-03-16. 2

[8] An incredible series of videos swaps famous hollywood faces to demonstrate how convincing 'deepfake' tech has gotten. https://www.businessinsider.com/deepfakes-of-famous-movies-youtube-channel-2019-5. Accessed: 2020-05-27. 2

[9] Pai synthetic media framework. https://syntheticmedia.partnershiponai.org/. Accessed: 2023-03-05. 1

[10] vit-pytorch. https://github.com/lucidrains/vit-pytorch. Accessed: 2022-11-12. 5, 7

[11] 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. 2

[12] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2814–2822, 2020. 2

[13] Michael Albright and Scott McCloskey. Source generator attribution via inversion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[14] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1205–1207, 2019. 2

[15] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. 2021. 3

[16] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 5

[17] 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. 2

[18] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 5

[19] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, Los Alamitos, CA, USA, may 2017. IEEE Computer Society. 2

[20] Umur Aybars Çiftçi, İlke Demir, and Lijun Yin. Fake-Catcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis & Machine Intelligence (PAMI)*, 2020. 2, 3, 6

[21] Umur Aybars Çiftçi, İlke Demir, and Lijun Yin. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020. 3, 6, 8

[22] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[23] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 6

[24] David Chu, İlke Demir, Kristen Eichensehr, Jacob G Foster, Mark L Green, Kristina Lerman, Filippo Menczer, Cailin O'Connor, Edward Parson, Lars 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. 1, 2

[25] Umur Aybars Çiftçi, İlke Demir, and Lijun Yin. Deepfake source detection in a heart beat. *The Visual Computer*, pages 1–18, 2023. 6

[26] M. Civera, L. Zanotti Fragonara, and C. Surace. An experimental study of the feasibility of phase-based video magnification for damage detection and localisation in operational deflection shapes. *Strain*, 56(1):e12336, 2020. e12336 STRAIN-1499.R1. 3

[27] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, pages 219–229, Cham, 2022. Springer International Publishing. 2

[28] Rashmiranjan Das, Gaurav Negi, and Alan F. Smeaton. Detecting deepfake videos using euler video magnification. *Electronic Imaging*, 33(4):272–1–272–7, jan 2021. 4, 6

[29] İlke Demir and Umur A. Çiftçi. MixSyn: Learning composition and style for multi-source image synthesis. *CoRR*, abs/2111.12705, 2021. 2

[30] İlke Demir and Umur Aybars Çiftçi. Where do deep fakes look? synthetic face detection via gaze tracking. In *ACM Symposium on Eye Tracking Research and Applications*, New York, NY, USA, 2021. Association for Computing Machinery. 2, 8

[31] Yuzhen Ding, Nupur Thakur, and Baoxin Li. Does a gan leave distinct model-specific fingerprints? In *BMVC*, 2021. 3

[32] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. 2019. 3

[33] Jianwei Fei, Zhihua Xia, Peipeng Yu, and Fengjun Xiao. Exposing ai-generated videos with motion magnification. *Multimedia Tools Appl.*, 80(20):30789–30802, aug 2021. 4

[34] Camilo Fosco, Emilie Josephs, Alex Andonian, and Aude Oliva. Deepfake caricatures: Human-guided motion magnification improves deepfake detection by humans and machines. *Journal of Vision*, 22(14):4079–4079, 2022. 2

[35] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 1, 2

[36] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[37] 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. 2

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[39] Brian Hosler, Davide Salvi, Anthony Murray, Fabio Antonacci, Paolo Bestagini, Stefano Tubaro, and Matthew C. Stamm. Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1013–1022, June 2021. 8

[40] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2500–2504, 2021. 2

[41] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. 6

[42] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv:1602.07360*, 2016. 7

[43] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018. 2

[44] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2895, 2020. 3

[45] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2017. 2

[46] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[47] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021. 2, 3, 5

[48] 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. 2

[49] Okan Köpüklü, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1910–1919. IEEE, 2019. 5

[50] P. Korshunov and S. Marcel. Speaker inconsistency detection in tampered video. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2375–2379, Sep. 2018. 2

[51] Alicja Kwasniewska, Jacek Ruminski, and Maciej Szankin. Improving accuracy of contactless respiratory rate estimation by enhancing thermal sequences with deep neural networks. *Applied Sciences*, 9(20), 2019. 3

[52] Nam Le and Jean-Marc 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. 3

[53] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2

[54] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009, 2020. 2

[55] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. 2

[56] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020. 3

[57] Michael Lomnitz, Zigfried Hampel-Arias, Vishal Sandesara, and Simon Hu. Multimodal approach for deepfake detection. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9, 2020. 2

[58] 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. 3

[59] 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. 3

[60] Aman Mehra, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Motion magnified 3-d residual-in-dense network for deepfake detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1):39–52, 2022. 2, 4, 6

[61] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1), jan 2021. 2

[62] Ronald Mutegeki and Dong Seog Han. A cnn-lstm approach to human activity recognition. In *2020 international conference on artificial intelligence in information and communication (ICAIIC)*, pages 362–366. IEEE, 2020. 7

[63] Pedro C Neto, Eduarda Caldeira, Jaime S Cardoso, and Ana F Sequeira. Compressed models decompress race biases: What quantized models forget for fair face recognition. *arXiv preprint arXiv:2308.11840*, 2023. 7

[64] 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. 2

[65] Ewa M. Nowara, Daniel McDuff, and Ashok Veeraraghavan. Combining magnification and measurement for non-contact cardiac monitoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3810–3819, June 2021. 3

[66] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Frédo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. *arXiv preprint arXiv:1804.02684*, 2018. 2, 4

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[68] KR Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 2

[69] Jiameng Pu, Neal Mangaokar, Lauren Kelly, Parantapa Bhattacharya, Kavya Sundaram, Mobin Javed, Bolun Wang, and Bimal Viswanath. Deepfake videos in the wild: Analysis and detection. 2021. 3

[70] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. *arXiv e-prints*, page arXiv:1803.09179, Mar 2018. 3

[71] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 5

[72] Sophie R. Saremsky, Umur A. Ciftci, Emily A. Greene, and Ilke Demir. Adversarial deepfake generation for detector misclassification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[73] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 6

[74] Jeremy Straub. Using subject face brightness assessment to detect 'deep fakes' (Conference Presentation). In Nasser Kehtarnavaz and Matthias F. Carlsohn, editors, *Real-Time Image Processing and Deep Learning 2019*, volume 10996. International Society for Optics and Photonics, 2019. 2

[75] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 6

[76] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon 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. 2

[77] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), July 2019. 2, 4

[78] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6), Oct. 2015. 2

[79] 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. 2

[80] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 2

[81] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 5, 7

[82] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Phase-based video motion processing. *ACM Trans. Graph.*, 32(4), jul 2013. 2, 5

[83] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 3

[84] Zhaoqiang Xia, Xiaopeng Hong, Xingyu Gao, Xiaoyi Feng, and Guoying Zhao. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Transactions on Multimedia*, 22(3):626–640, 2020. 3

[85] 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. 2, 3

[86] Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[87] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[88] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 7

[89] 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. 2

[90] 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. 2