Generalizable Deepfake Detection With Phase-Based Motion Analysis

Ekta Prashnani, Michael Goebel, and B. S. Manjunath, Fellow, IEEE

Abstract—We propose PhaseForensics, a DeepFake (DF) video detection method that uses a phase-based motion representation of facial temporal dynamics. Existing methods that rely on temporal information across video frames for DF detection have many advantages over the methods that only utilize the per-frame features. However, these temporal DF detection methods still show limited cross-dataset generalization and robustness to common distortions due to factors such as error-prone motion estimation, inaccurate landmark tracking, or the susceptibility of the pixel intensity-based features to adversarial distortions and estimation, inaccurate landmark tracking, or the susceptibility of common distortions due to factors such as error-prone motion estimation. Our key insight to overcome these issues is to leverage the temporal phase variations in the band-pass frequency components of a face region across video frames. This not only enables a robust estimate of the temporal dynamics in the facial regions, but is also less prone to cross-dataset variations. Furthermore, we show that the band-pass filters used to compute the local per-frame phase form an effective defense against the perturbations commonly seen in gradient-based adversarial attacks. Overall, with PhaseForensics, we show improved distortion and adversarial robustness, and state-of-the-art cross-dataset generalization, with 92.4% video-level AUC on the challenging CelebDFv2 benchmark (a recent state-of-the-art method, FTCN, compares at 86.9%).

Index Terms—Deepfake detection, deep learning, video analysis, video forensics.

I. INTRODUCTION

Deep generative models for faces [44], [45], with their ability to create realistic-looking facial animations, have enabled fields like creative communication, and film-making. Their increasing efficacy also calls into consideration the impact of such technology on society, if misused to spread false information [24]. This has led to an active area of research – deepfake (DF) detection [9], [71], [82] – which aims to tell apart an original face image/video from a “fake” one created using a DF generator. Broadly, video DF detection methods either rely on the spatial per-frame artifacts left by DF generators (such as warping [51] or upsampling [54]), or the anomalies in the facial temporal dynamics of a DF video (such as mouth movements [27], [35]). While the per-frame spatial artifacts are easier to overcome with the recent advancements in the face image generators [44], the temporal inconsistencies of such generators are still evident in the generated DF videos [35]. DF detectors relying on temporal cues are therefore more effective.

We contribute to the promising trend – DF detection based on temporal dynamics – with our proposed method, PhaseForensics (Fig. 1), that leverages temporal phase changes to estimate a noise-robust, domain-invariant, representation of the facial motion. By relying on the link between local temporal phase changes and local motion in a video (per Fourier Shift theorem), we learn a DF detector directly from the facial temporal dynamics instead of the per-frame artifacts. As compared to existing methods based on frequency/phase domain – that use phase and frequency-based features to isolate spatial cues such as per-frame upsampling [54], or high-frequency artifacts [56] – our method leverages a fundamentally different property for DF detection: the local motion in a video, as estimated by the local temporal phase variations.

PhaseForensics offers three important advantages over existing temporal DF detectors. First, by relying on the temporal phase changes across frames as a measure of the facial dynamics [28], we avoid the error-prone approach of explicitly estimating facial motion. Some of the existing temporal DF detectors that rely on facial dynamics leverage tracking hand-coded features across frames (e.g., motion vectors [5]), landmark trajectories [73], or temporal dynamics learnt from RGB inputs [35]: this makes such methods prone to inaccuracies due to tracking errors. PhaseForensics does not depend on tracking/motion estimation, but still provides an indication of the amount of motion in face sub-regions.

The second advantage of the phase domain is that it affords increased robustness to appearance changes (e.g., contrast or scale changes) [28], [31]. Consequently, we observe improved robustness to spatial distortions associated with color, noise, and compression artifacts, as well as state-of-the-art cross-dataset generalization (e.g., with an AUC of 92.4% on the challenging CelebDFv2 dataset [52]). In Fig. 1b, we demon-
We will now discuss the existing notable works for DF detection. In our discussion, we classify the existing methods based on their different design choices, with some overlap across different categories. We refer the readers to comprehensive reviews for a more in-depth consideration [9], [82], including those that focus on leveraging deep learning-based models for DF detection [64], [68].

A. Frame-Based Methods

The sub-field of frame-based DF detection has witnessed active progress over the years, with early methods performing per-frame feature-extraction using pretrained networks [69] to more sophisticated frame-based approaches utilizing attention or contrastive learning [72]. For example, [84] performs fine-grained per-frame classification using the learnt multi-attentional maps with soft-attention dropout and a regional independence loss, [78] proposes an image-based method that uses multi-scale transformers to process RGB image, followed by fusion of processed features with the DCT domain, and [72] leverages contrastive learning to detect real vs. fake images. Other works adopt techniques for interpretable DF detection using channel-wise correspondence [38], LSTMs, or transformers [80], or coordination across facial features [12]. The typical loss terms used for training are a combination of cross-entropy, segmentation, and contrastive loss. While frame-based methods have been popular, dependence on single-frame DF generation artifacts can be prone to inaccuracies. This is because of the rising sophistication of DF generators [44], [45], and also because the imprints left by such generators can easily be lost by simple architecture changes [11]. This can especially impact methods that rely on signals such as generative model footprint [54], face warping artifacts [51], or blending boundaries for classifying real vs. fake faces [50]. Therefore, several other methods have attempted to improve detection performance and generalization by leveraging additional features. Such features have included Discrete Cosine Transform (DCT) coefficients [67], mid-level features [4], and Laplacian of Gaussians [59].

B. Methods Relying on Biological or Identity-Based Features

Many methods rely on detecting inconsistencies in facial landmarks or identity-specific behaviors between real and deepfake videos [20], [22], [26], [34], [73]. Some of these methods leverage existing, pretrained face landmark detection models, to extract a condensed feature to be passed for classification. Others estimate identity-specific behaviors using facial geometry changes [20], [26] or typical statistics in real videos [34], followed by estimating deviations from these to detect deepfake videos. While our work focuses purely on video modality, a related class of person-specific DF detection methods are also multi-modal: these focus on audio-visual consistencies and person-specific idiosyncrasies for DF detection [19]. While these methods may offer a degree of explainability not obvious in more complicated neural networks, the performance results generally lag behind...
the end-to-end trainable models. Related to these are the methods that rely on biological signals – with the aim to track inconsistencies in heart-rate [17], [18], [21], [37] or blink/gaze patterns [18], [22], [43]). With these approaches, there is a susceptibility towards errors since such methods depend on reliable estimation of the biological signals – which can be lost in low-quality videos or inevitable changes in person-specific facial behaviors with time or evolving context.

C. Methods Leveraging Phase/Frequency Information

An important sub-class of DF detection methods have considered leveraging frequency and phase-related cues. Specifically, these include phase or frequency spectrum imprint of generative models [30], [47], [54], [56], per-frame Laplacian of Gaussians [57], tracking frequency-level perturbations from the DF generation process [41], variations in wavelet-based statistics [48], and DCT coefficients [67]. All these methods use the frequency or phase domain to estimate frame-level features to inform DF detection. In contrast, our novelty lies in effectively leveraging temporal phase variations to estimate a domain-invariant representation of local motion in face regions. Therefore, we leverage a fundamentally different aspect for DF detection compared to these methods.

D. Methods That Perform Temporal Processing

In order to overcome the potential limitations of frame-based approaches, a recent trend has been to perform sequence modeling – to make the DF detector dependent on temporal signals. Some of the existing methods estimate per-frame features that are fed into Recurrent Neural Networks (RNNs) [33], [36], [39], [70] or Long-Short Term Memory (LSTM) [13], [15] networks, and trained end-to-end [33], [36], [39], [70]. Reference [59] present a two-branch RNN-based spatio-temporal deepfake detector, utilizing Laplacian of Gaussian to amplify medium and high-frequency artifacts, while the second branch uses RGB data. Reference [73] model the temporal behavior of landmarks by providing relevant features to a two-stream RNN for classification. Reference [55] learn spatial and temporal attention maps used to modulate shallow and mid-level features with the aim of capturing long-spatial-distance dependencies and anomalies in the coordination of facial features. Reference [83] leverage 3DCNN-based features with temporal dropouts applied during training for effective DF detection. Reference [32] leverage latent sensing for video-based DF detection. A closely related work to ours is LipForensics [35]: a multi-scale temporal CNN is used to classify deepfakes based on temporal features extracted from lip region using a ResNet-18 and 3DCNN layers. As we will elaborate in the later sections, our choice of phase-based temporally-processed input features (as opposed to pixel intensity input from LipForensics) and the choice temporal processing modules enables far better cross-dataset generalization (92.4% AUC on the challenging CelebDFv2 dataset [52]; vs. 82.4% from LipForensics) and improved robustness to spatial and adversarial distortions. Consistent with our own observations, the fact that mouth region provides revealing cues for DF detection based on facial dynamics has also been discussed in previous works [27], [46].

E. Choice of Pretraining Datasets

A noteworthy design strategy with many DF detection methods, across all categories discussed above, has been to pretrain on suitable datasets. While some works adopt domain-specific training strategy [35], others adopt more generic pretrained networks, such as those that are trained on ImageNet [69].

Despite these recent advances, we observe that the generalization and robustness of temporal DF detectors can be improved. To achieve this with PhaseForensics, we move away from depending on pixel-intensity-based features, and instead rely on temporal phase variations robustly isolate the facial dynamics for DF detection.

III. PHASEFORENSICS: PHASE-BASED DF DETECTION

For any given input video, we first apply a standard pre-processing pipeline to detect the face region in the input video, stabilize it, and crop out the relevant portion. After this, we apply the two-step PhaseForensics pipeline (Fig. 1a). First, we estimate the local per-frame phase from spatial sub-bands (using the complex steerable pyramid [65]), which we then pass to a learnable spatio-temporal filter to isolate the relevant facial dynamics from phase for DF detection. Second, we use these spatio-temporally filtered phase-based features as input to train a standard DF detection pipeline comprising of a feature extractor and a multi-scale temporal convolutional network that is known to effectively perform sequence modeling [35], [53], [58]. In the next section, we begin the discussion of our method by first elaborating upon the intuition behind using phase-based inputs, as compared to RGB-domain inputs.

A. Intuition Behind Using Phase Instead of Pixel Intensity

Our method relies on the fact that realistically synthesizing facial movements is tough: subtle inconsistencies in facial motion are there in even hyper-realistic DF videos. Most existing temporal methods use color or pixel-intensity domain input that fails to capture these subtleties, since motion and appearance cues are coupled in these domains (e.g., motion artifacts may not be evident in low-contrast face areas or when the motion amplitude is too small). Our insight is that these subtle inconsistencies are more evident in the phase domain since the phase variations solely depend on motion, independent of the amplitude of the original signal. In Fig. 3, we show an example of how phase effectively identifies inconsistencies in real vs. DF videos, by comparing the temporal phase and color variations at facial landmarks for a real and a DF video from VHQ [29]. The DF video is a realistic-looking reconstruction of the real one (with no apparent visual artifacts) – obtained by animating the first frame of the real video with expressions from subsequent frames of the same video. This procedure for reconstruction means that facial expressions in both real and fake video frames should correspond (ideally – in case of a perfect generation process). For both videos, from left to right, we show: the video frames with the facial landmark annotations, and the phase and color values (RGB) plotted as a
function of time at a landmark on the lip. It is evident that the temporal dynamics of phase show clear differences between the real and the fake video, while such differences are not apparent in the temporal RGB plots.

This is further confirmed by the Pearson’s linear correlation (corr.) between the two phase plots which is 0.32, while for the RGB plots it is 0.9. We also reached the same conclusion for temporal phase and color variations at other landmarks (annotated on the frames; corr. shown in the table). This confirms our intuition temporal phase variations better isolate inconsistencies in facial motion of deepfakes.

Next, we discuss the details of our method, starting with spatio-temporal feature estimation.

B. Spatio-Temporal Feature Estimation

For an image, \( I(x, y) \), global translations along \( x \) and \( y \) directions, such as \( I(x + \Delta x, y + \Delta y) \), directly relate to the phase changes in the Fourier coefficients of \( I(x, y) \). Specifically, for all frequency components \((\omega_x, \omega_y)\), the phase change due to the translation would correspond to \((\omega_x \Delta x, \omega_y \Delta y)\) as per the Fourier Shift Theorem. Given their infinite spatial support, phase changes in the Fourier basis functions directly capture global translation in \( I(x, y) \). Intuitively, this idea extends to local spatial shifts: looking at the phase changes in an image decomposition constituted by finite-support quadrature filter pairs – such as with complex wavelets, or, its polar-separable version, the complex steerable pyramid [65] – provide an indication of the local shift in image content. In a typical video, \( V(t, x, y) \), (e.g. that of a talking face), motion is spatially localized, and can be captured with such local phase changes, more effectively than the phase variations in Fourier coefficients – a phenomenon we also demonstrate later through experiments. This facial motion may differ across multiple scales, and may not be effectively captured by a single motion estimation filter. To represent these multiple granularities of motion, we utilize the complex steerable pyramid (CSP) decomposition [65] and compute the phase of the complex coefficients obtained from the scaled and oriented bandpass components of the per-frame CSP. Temporal changes in the phase of these components represents local motion in the video at the different orientations and scales. The low-pass and high-pass residual components of CSP, being real-valued, do not contain phase information. We therefore only consider the bandpass components of the CSP in PhaseForensics. Moreover, since these residual components directly relate to image intensities, they are prone to spatial perturbations and domain-specific cues. We justify our choice of discarding these residual components in Sec. IV, where we compare the result of using CSP residuals for training as well. As an aside, we note that CSP is also a popular and effective choice for estimating local motion for other video-related tasks such as motion magnification [76], motion interpolation [60], and motion transfer [66].

For the input video frame at time instance \( t \), \( V(t, x, y) \), the complex-valued coefficient, \( R_{\omega,\theta}(t, x, y) \), obtained after filtering with the spatial bandpass filter, \( \Psi_{\omega,\theta}(x, y) \) [65], at scale \( \omega \) and orientation \( \theta \) is computed as:

\[
R_{\omega,\theta}(t, x, y) = V(t, x, y) * \Psi_{\omega,\theta}(x, y)
= A_{\omega,\theta}(t, x, y)(C_{\omega,\theta}(t, x, y)
+ i S_{\omega,\theta}(t, x, y)).
\]

Here, \( A_{\omega,\theta}(\cdot) \) is the amplitude of the complex response, and \( A_{\omega,\theta}(\cdot)C_{\omega,\theta}(\cdot), A_{\omega,\theta}(\cdot)S_{\omega,\theta}(\cdot) \) denote the responses to the quadrature filter pairs, with phase \( \Phi_{\omega,\theta}(t, x, y) = \arctan(S_{\omega,\theta}(t, x, y)/C_{\omega,\theta}(t, x, y)) \). The phase estimates from all scales and orientations of the CSP are concatenated along a new dimension, which we term \( c \), to yield a per-frame tensor at each instance \( t \) which we denote as \( \Phi(t, x, y, c) \).

Given \( \Phi(t, x, y, c) \) for all the video frames, we now want to isolate phase variations across time that are relevant to DF detection. Traditionally, the motion of relevant objects is isolated by a hard-coded temporal filter \( \Phi(t, x, y, c) \) [62]. However, hardcoding a temporal filter design can be suboptimal for DF detection: for example, when capturing motion of lip region, it is hard to guarantee that a hardcoded temporal filter would capture lip movements at different talking speeds. Instead, we allow our DF detector to guide this process of isolating the temporal phase changes relevant to DF detection through an end-to-end learning process. Therefore, we apply a learnable temporal filtering operation \( f_t \) (with parameters \( \theta_t \)) to \( \Phi(t, x, y, c) \). Before temporal filtering, we perform a spatial filtering operation (again, learnable – with parameters \( \theta_s \)) \( s_{x,y} \) on \( \Phi(t, x, y, c) \), to overcome any spurious spatial artifacts and improve the signal-to-noise ratio of \( \Phi(t, x, y, c) \) (previous work do this with a hard-coded filter [76]). A noteworthy issue with \( \Phi(t, x, y, c) \) is the ambiguity around large motions since phase shifts beyond \( \pm \pi \) are ill-defined. One way to resolve this to some extent is explored in motion interpolation works [60] – where phase information of different scales are used to resolve the ambiguity for a given scale. With our proposed construction of \( \Phi(t, x, y, c) \), our deep learning model has the flexibility to also perform this operation. This is because the sub-bands across multiple scales in \( \Phi(t, x, y, c) \) are stacked along the channel dimension – allowing for across-scale interactions in a typical conv layer. Specifically, in the implementation of the 3D separable convolution, the receptive field of \( f_t \) and \( s_{x,y} \) along the channel dimension ensures this interaction along the various sub-bands, which can help resolve the ambiguity. In summary, the post-processing operations on \( \Phi(t, x, y, c) \) are implemented as a learnable, separable, 3D convolution (as done in [81]), to yield,

\[
\Phi^p(t, x, y, c) = f_t(s_{x,y}(\Phi(t, x, y, c))),
\]

which we use for our DF detection (super-script \( p \) indicates processed output). By design, phase can only be extracted from the bandpass components of the CSP. Therefore, we use only the bandpass components. For completeness, we evaluate alternative choices such as training using the entire set of CSP coefficients and other frequency decompositions as well.

**Alternatives to CSP:** An alternative to obtain phase information is discrete Fourier transform (DFT): as discussed earlier, the infinite spatial support of the DFT basis makes it impossible to use its coefficients to estimate of local motion,
while the finite spatial support of the CSP basis allows for this. Other sub-band decompositions can include wavelets or Laplacian of Gaussians [59], both of which, being real-valued, do not provide local phase information. This makes CSP the best choice amongst alternatives.

C. Implementation Details

1) Facial Region Selection: Before training with $\Phi^p(t, x, y, c)$ for DF detection, we want to understand which facial regions can benefit the most from such an approach. This is because, for regions that contain very little or no motion, variations in $\Phi^p(t, x, y, c)$ are meaningless [76], and can adversely affect the training process. Since the lip region of the face provides the largest motion cues, for our main experiments we only focus on the lip region of the face to compute $\Phi^p(t, x, y, c)$, similar to a previous work, LipForensics [35]. In Sec. IV-B, we also apply PhaseForensics to the eye and face regions, and compare it to our experiments with the lip region. Our process of facial sub-region extraction and alignment follow a standard pipeline – with face detection, landmark detection, and alignment, followed by cropping, consistent with existing works [34], [35]. Specifically, we first detect faces in each frame [23]. The frames are then cropped to the largest facial bounding box and resized to $256 \times 256$, followed by landmark detection [7], and alignment of frames to rigid landmarks of an average face [35]. Note that this removes the headpose changes, but does not affect the non-rigid facial motion (expression, lip movements), which is what we want to isolate for DF detection. We centrally crop facial regions from the aligned frames as follows: 1) for lips, we crop an $88 \times 88$ region around lips, 2) $128 \times 128$ region for the models that use the full face, and 3) $128 \times 88$ around eyes for the models that use only the eye region. During training, as a way to regularize the training process, we first increase the crop size by 10% and randomly crop facial regions to the aforementioned sizes.

2) Architecture: For a given video clip, the spatio-temporally processed phase, $\Phi^p(t, x, y, c)$ (Eq. 2) is passed to the next stage of the DF detection pipeline. We adopt a standard two-step approach for the rest of our DF detection pipeline: we first compute per-frame feature embeddings, followed by sequence modeling to learn the temporal behavior (Fig. 2). We choose to use the ResNet-18 architecture (with modified input channels to match the maximum value of $\Phi^p(t, x, y, c)$) as our feature extractor – denoting this function as $r_{x,y}$, with learnable parameters $\theta_r$. For temporal modeling, we choose a multi-scale temporal convolutional network (MSTCN - same as [6]), given their remarkable performance gain over typically-used LSTMs for a variety of tasks [35], [53], [58] and due to their flexible design with varying temporal receptive field and lightweight construction. The output of MSTCN is passed through a linear layer to yield the final prediction. We denote the function represented by MSTCN + linear classifier as $m_{r_{x,y}}$, with learnable parameters $\theta_m$. The output logit from PhaseForensics pipeline for a given input video clip $V(t, x, y)$, is then given by $\hat{y} = m_{r_{x,y}}(\Phi^p(t, x, y, c))$. We train PhaseForensics in an end-to-end manner to optimize all parameters, $\theta_f, \theta_s, \theta_r, \theta_m$ by minimizing the average binary cross-entropy loss across training samples $\frac{1}{N} \sum_{i=1}^{N} L_{BCE}(\hat{y}_i, y_i^f, \theta_f, \theta_s, \theta_r, \theta_m)$, where $i$ is the training sample and $N$ is the training dataset size.

3) Pretraining and Hyperparameters: Before training PhaseForensics for DF detection, we perform domain-specific pretraining on the lip and eye regions, with $\Phi^p(t, x, y, c)$ network.
(Eq. 2) as input. We pretrain the architecture discussed above on tasks relevant to these facial sub-regions, since these are shown to significantly improve the generalization of the DF detector [35]. We note that many competing methods adopt unique training steps best-suited for their specific goals, such as pretraining on lipreading [35], ImageNet pretraining [59], [69], or using custom dataset [52]. In our evaluation of competing methods, we regard the author-prescribed training steps as the best practice – without taking away any of the key steps. This is also commonly done in performance comparisons presented by existing state-of-the-art DF detection methods [35], [85]. Similar to LipForensics [35] (a state-of-the-art method leveraging anomaly detection), training PhaseForensics on, say, lips, involves first learning the distribution of natural temporal dynamics of lips using the Lip Reading in the Wild (LRW) dataset [16] for 50,000 iterations, with a cross-entropy loss (similar to [35]), and then finetuning for DF detection. More formally, there are two steps: 1) learning the natural lip movements by training for lipreading, 2) learning to detect anomalies in the lip movements of deepfakes by training on DF dataset. In Sec. IV, we clearly motivate this two-step approach by demonstrating the drop in performance observed when LRW pretraining is not performed. Similarly for the eye region, we pretrain for gaze prediction on the EVE dataset [63] for 50,000 iterations, with crops from face images showing both the eyes, and an angular loss [63], that measures the error between predicted and true gaze. This leads to an accuracy of at least 80% for all pretrained models, which provides sufficient domain-specific knowledge to then fine-tune these models for DF detection. Based on our experiments, training solely on lip region yields the best DF detection performance. Moreover, there are many configurations we could adopt for computing \( \Phi(t, x, y, c) \) from the CSP decomposition, by varying the number of scales, orientations, and bandwidths of the filters. Our best-performing choice is to compute \( \Phi(t, x, y, c) \) using 4 spatial scales, octave bandwidth, and 2 orientations for lip sub-images (this CSP configuration is represented as “4s2o-oct”), and 3 spatial scales, octave bandwidth, and 4 orientations for eye region. For the models trained on lip regions, we also demonstrate the effect of using alternative CSP decompositions (varying number of orientations, and filter bandwidth) to verify the best choice. Lastly, for both the training stages, Adam optimizer with a learning rate of \( 2e^{-4} \) and a batch size of 32 is used. The duration of an input clip is 25 frames, and we sample all clips exhaustively from a video. The DF detection training is stopped when the validation loss does not show any improvement for 20,000 training iterations.

IV. RESULTS

We now compare PhaseForensics to existing popular and state-of-the-art DF detection methods, and motivate our design choices with ablation studies. Fig. 5 shows an overview of our comparisons against existing methods: we evaluate cross-dataset generalization, and, in case temporal of DF detection methods, robustness to spatial and adversarial distortions. We consider the classic frame-based methods such as the Xception baseline [69]; recent popular approaches such as PatchForensics [10] (truncated Xception classifier trained on aligned faces, with result averaged over patches [35]), Multi-Attention [84], Face X-ray [52] (from [35], trained with blended images and fake samples), DSP-FWA [51], local relation-based method [12], patch-channel correspondence method [38], contrastively-trained models [72]; and temporal methods such as CNN-GRU [70] (DenseNet-161 [39] trained with GRU [114]), 3DCNN [83], LipForensics [35], FTCN [85] a video-representation learning method [32], and TD-3DCNN [83]. We further emphasize that some of the methods we compare against specifically leverage features derived from the frequency domain. These methods include traditional approaches such as Two-branch [59], as well as recent methods that leverage phase-based discrepancies such as SPSL [54], and those that leverage wavelet-based features [48]. We analyze DFT-based inputs in Sec. IV-B. For PhaseForensics, our default, best-performing, model is obtained by training on only the lip region, with 4-scale, 2-orientation, and octave-bandwidth filters in the CSP (this CSP configuration is represented as “4s2o-oct”). Given that our best model works with the lip region, we consider LipForensics as an important baseline in our evaluations [35]. In our ablation studies, we compare against LipForensics to demonstrate the advantage of using phase-based features over pixel intensity-based input used in LipForensics. Consistent with recent approaches [35], [69], we report the video-level Area Under ROC Curve (AUC; the result is averaged over frames for frame-based methods).

A. Evaluating the Generalizability of DF Detectors

Recent DF detection methods tend to show near-perfect performance when evaluated on within-domain videos (i.e., test set of FF++) – which does not give a clear sense of real-world generalizability of such methods. Our main aim in this section is, therefore, to thoroughly analyze the generalization to completely new datasets never seen during training. Moreover, for temporal DF detection methods, we also analyze their robustness to spatial and adversarial distortions, also within the context of cross-dataset generalizability.

1) Training Dataset: PhaseForensics and all other methods evaluated here are trained on FaceForensics++ (FF++) training set [69]. FF++ comprises of a total of 1000 unmanipulated videos, and corresponding manipulated videos with 4 DF generation methods – 2 each for face-swapping (DeepFakes [2], FaceSwap [3]) and face re-enactment (Face2Face [75], NeuralTextures [74]) DF generation approaches. We adopt the train/val/test splits specified by the dataset and use the trained model from FF++ training for all experiments.
TABLE I
HERE WE SHOW THE VIDEO-LEVEL AUC (%) FOR ALL MODELS TRAINED WITH FACEFORESICS++ (FF++) [69]. SOME NUMBERS ARE REPORTED FROM EXISTING BENCHMARKS [35], [52]. WHILE A ‘−’ IS STATED WHEN THE METRIC IS NOT AVAILABLE (MISSING CODE/TRAINED MODEL). (A) SHOWS THE CROSS-DATASET GENERALIZATION ON DFD [1], DFDCC [25], VFHQ [29], AND CDFV2 [52] DATASETS, AND (B) EVALUATES THE TRAILED MODELS ON FF++ TEST VIDEOS, AND ALSO WITH DIFFERENT DF GENERATORS USED TO MANIPULATE ORIGINAL FF++ VIDEOS. PHASEFORESICS ACHIEVES STATE-OF-THE-ART RESULTS (SHOWN IN BOLD) FOR CROSS-DATASET GENERALIZATION, AS COMPARED TO EXISTING METHODS. IN (B), PHASEFORESICS RANKS AMONGST THE TOP METHODS (TOP-3 ARE UNDERLINED) WHEN TESTED ON NOVEL (NOT SEEN IN TRAINING) DF GENERATION METHODS APPLIED TO THE ORIGINAL FF++ TEST VIDEOS (FSh [49] AND DFor [42]). UNLIKE PHASEFORESICS, MOST METHODS ACHIEVE HIGH ACCURACY ON THE WITHIN-DOMAIN FF++ TEST SET IN (B), WHILE GENERALIZING POORLY TO NEW DATASETS IN (A).

| METHOD                   | DFDCC | VFHQ | CDFV2 | DFD |
|--------------------------|-------|------|-------|-----|
| Xception [69]            | 70.9% | 70.1%| 73.7% | -   |
| Multi-Attention [84]     | 63.0% | 55.0%| 68.0% | -   |
| PatchForensics [10]      | 65.6% | -    | 69.6% | -   |
| Face X-ray [59]          | 65.5% | -    | 79.5% | -   |
| CNN GRU [70]             | 68.9% | 66.0%| 69.8% | 70.2% |
| Two-branch [59]          | -     | -    | 76.7% | -   |
| DSP FWA [51]             | 67.3% | 69.0%| 69.5% | -   |
| SPL [54]                 | 66.2% | -    | 76.88%| -   |
| TD-3DCNN [53]            | 55.0% | -    | 57.3% | -   |
| Local-relation [12]      | 76.5% | -    | 78.3% | 89.2% |
| Wavelet-enhanced [48]    | -     | -    | 84.8% | -   |
| Patch-corresp. [35]      | 62.7% | -    | 54.9% | -   |
| DCL [72]                 | 76.7% | -    | 82.3% | 91.7% |
| video rep. learning [32] | 73.2% | -    | 87.7% | -   |
| LipForensics [35]        | 73.5% | 90.2%| 82.4% | 91.9% |
| FTCN [85]                | 74.0% | 84.8%| 86.9% | -   |
| PhaseForensics (4s2o-oct)| 76.9% | 92.3%| 92.4% | 92.1% |

(a) cross-dataset generalization

| METHOD                   | FF++ | DFor | FSh |
|--------------------------|------|------|-----|
| Xception [69]            | 99.8%| 84.5%| 72.0%|
| Multi-Attention [84]     | 89.8%| 72.3%| 60.2%|
| PatchForensics [10]      | 99.9%| 81.8%| 57.8%|
| Face X-ray [50]          | 99.8%| 86.8%| 92.8%|
| CNN GRU [70]             | 99.9%| 74.1%| 80.8%|
| Two-branch [59]          | 99.1%| -    | -   |
| DSP FWA [51]             | 75.7%| 50.2%| 65.5%|
| SPL [54]                 | 95.32%| -    | -   |
| TD-3DCNN [83]            | 72.22%| -    | -   |
| Local-relation [12]      | 97.6%| -    | -   |
| wavelet-enhanced [48]    | 99.6%| -    | -   |
| Patch-corresp. [38]      | 99.8%| -    | -   |
| DCL [72]                 | 99.3%| -    | -   |
| video rep. learning [32] | 99.9%| -    | -   |
| LipForensics [35]        | 99.9%| 97.6%| 97.1%|
| FTCN [85]                | 99.7%| 98.8%| 98.8%|
| PhaseForensics (4s2o-oct)| 99.6%| 93.6%| 95.6%|

(b) cross-manipulation generalization

2) Evaluation Approach, Datasets, and Metrics: For cross-dataset generalization test, we use three datasets: CelebDFv2 (CDFv2) – containing very high-quality face-swapped deepfakes [52]; the recent DeepFakeDetection (DFD) dataset [1]; DFDCC test subset – featuring face-swapping and face re-enactment deepfakes [25], and the VideoForensicsHQ dataset – a very high-quality face re-enactment DF dataset (VFHQ) [29], as shown in Tab. Ia. We evaluate all models on FF++ test videos with the same face manipulations as seen in training (i.e., within-domain performance), and on novel face manipulation methods applied to the FF++ videos (i.e., cross-manipulation performance), such as DeeperForensics (DFor) [42] (without spatial distortions), and FaceShifter (FSh) [49]. We note that, existing works have shown this cross-manipulation test by training on 3 of the 4 facial manipulations in FF++ dataset and evaluating on the remaining one. In contrast, we test for this for cross-manipulation generalization in Tab. Ib using more recent DF generation methods that are also more challenging due to fewer generation artifacts [49]. Lastly, we also analyze the robustness to common spatial distortions and adversarial perturbations in this cross-dataset evaluation setting using CDFv2, for temporal DF detectors (Tab. II, III).

3) Cross-Dataset Generalization: Given various data capture conditions and DF generator algorithms, the cross-dataset performance evaluation (Tab. Ia) on CDFv2, VFHQ, and DFDCC allows for a thorough analysis of the generalization capabilities of DF detectors. Due to the robustness to appearance changes (Sec. I), PhaseForensics (with setting “4s2o-oct”) yields state-of-the-art cross-dataset generalization, over competing temporal DF detectors such as LipForensics (AUC 82.4% on CDFv2) [35] and FTCN (AUC 86.9% on CDFv2) [85], with a performance of 92.4% on CDFv2, 92.3% on VFHQ, and 76.9% on DFDCC.

4) Evaluation on Different DF Generators: As mentioned earlier, the performance of recent DF detection methods is near-perfect (close to 99% AUC) when evaluated on FF++ test set videos (after training on FF++ train set) for most methods. This also holds true to some extent even for unseen DF generator algorithm, when the test videos being evaluated are from FF++. Tab. Ib shows the results for this cross-manipulation generalization analysis (last two columns). We note that PhaseForensics is amongst the top 3 methods in this analysis. We attribute the slight drop in cross-manipulation performance we observed with PhaseForensics to the tendency of our approach to learn facial motion representations that are not biased to the specific appearance/domain of the original videos in the FF++ dataset (which are in the test subset of all datasets for Tab. Ib). We emphasize that the cross-dataset tests are crucial for understanding the real-world generalizability of DF detectors. A near-perfect performance on the cross-manipulation tests (Tab. Ib) can be considered reliable when it also leads to improved cross-dataset generalization and distortion/adversarial robustness. The performance of PhaseForensics for all four cross-dataset tests (Tab. Ia) outweighs the small drop in performance for within-domain cross-manipulation results. PhaseForensics gives the best performance on the challenging cross-dataset tests and distortion/adversarial robustness, while being within 0.052 AUC of the best methods in cross-manipulation tests.
5) Robustness to Spatial Distortions (Tab. II): Color filters and compression are commonly applied to internet media. We explore how robust PhaseForensics is to changes in color saturation, contrast, resizing operation, and video compression (distortion levels are derived from DFor [42]), as compared to three popular state-of-the-art DF detectors: CNN-GRU [33], LipForensics [35], FTCN [85]. We chose color and compression-based distortions for this analysis since these occur ubiquitously in videos over the internet, and an effective DF detector needs to therefore be robust to these. For completeness, we analyze performance on less-common, more obvious distortions like block-like artifacts, noise, etc. in Sec. IV-B. Many previous works show the distortion robustness results on FF++ test set (the same DF generators are used for training) [35]. Here we analyze the distortion robustness on the test dataset of CDFv2, which is a more challenging real world use case, give the high quality DF generator and no overlap with training set. In (Tab. II), we report raw %AUC values for first two distortion levels – since those represent more realistic-looking artifacts one might encounter on the internet – and plot the %AUC as a function of distortion levels to show the performance with increasing distortion. A drop in performance is observed across all methods, with LipForensics showing a smaller drop compared to PhaseForensics for compression-related distortions. Overall, PhaseForensics is the best-performing method despite these distortions. Nevertheless, as also discussed in Sec. IV-B, future research on optimal frequency decompositions can further improve the robustness of phase-based approaches.

6) Effectiveness to Multiple Attacks/Spatial Distortions: While Tab. II demonstrates robustness of our method when only one kind of spatial distortion is applied, there is another question worth asking: would our method also be robust when multiple perturbations are applied to the videos? For example, it is common to see videos on the internet which have some color perturbation as well as compression artifacts. To assess the robustness to multiple attacks, we compare the performance of various temporal DF detection methods to PhaseForensics, when evaluated on videos containing a combination of color and compression-related distortions. Our experiments show that PhaseForensics is robust to multiple attacks, and compared favorably to some of the existing temporal DF detection methods (Tab. IV).

7) Adversarial Robustness: Recently, existing works have observed the vulnerability of DF detectors to black-box adversarial attacks [40]. Black-box attacks represent a plausible case where the detection model may be unknown to the adversary, but the adversary has the ability to test the model on a limited number of inputs. We assess the adversarial robustness

| Distortion type | level | CNN-GRU | LipForensics | FTCN | PhaseForensics |
|-----------------|-------|---------|--------------|------|----------------|
| none            | 0     | 69.8    | 82.4         | 86.9 | 92.4           |
| saturation      | 1     | 70.5    | 72.3         | 85.5 | 92.5           |
| contrast        | 2     | 69.4    | 72.2         | 85.6 | 92.4           |
| size            | 1     | 73.6    | 74.8         | 85.6 | 91.3           |
|                | 2     | 73.5    | 74.7         | 83.4 | 92.3           |
| compress        | 1     | 64.8    | 76.3         | 68.4 | 82.0           |
|                | 2     | 62.7    | 74.6         | 63.9 | 80.6           |

| TABLE III | EXAMPLES OF ADVERSARIAL PERTURBATIONS. Here We Show One Example of an Adversarial Perturbation Map for the Black-Box Attack [40] on LipForensics [35] and PhaseForensics (64x Amplified for Visualization). These Imperceptible Noise Signals Can Be Added to the Video, to Fool the DF Detection Method |

| model           | adv. robustness |
|-----------------|-----------------|
| CNN-GRU [70]    | 43.4            |
| LipForensics [35] | 49.5            |
| FTCN [85]       | 72.7            |
| PhaseForensics: 4x2o-oct | 76.5            |
| PhaseForensics: 4x4o-oct | 78.0            |
| PhaseForensics: 8x2o-half-oct | 81.0            |

input frame adv. perturbation (64x amplified for visibility) for PhaseForensics (left) and LipForensics (right)
for the existing temporal DF detectors – CNN-GRU [33], LipForensics [35], and FTCN [85] – and PhaseForensics with the default CSP configuration (4 scale, 2 orientations, octave bandwidth filters denoted as “4s2o-oct”) and two additional configurations for lip region: 1) 4 scales, 4 orientations, octave bandwidth filters denoted as “4s4o-oct”, and 2) 8 scales, 2 orientations, half-octave bandwidth filters denoted as “8s2o-half-oct”. We compute the adversarial perturbations for CDFv2 using a recent approach for black-box attacks for DF detectors (using the default settings provided by the authors) [40]. This black-box method utilizes Natural Evolutionary Strategies (NES) to estimate gradients between the input and output of the network without knowing the model weights. This is done by applying small perturbations to the input image in the form of Gaussian noise, and observing the change in output result. With an input image scaled to the range of \([0,1]\), we use 40 samples per step, and take a step size of 1/255. At most 25 total steps are taken, with a total perturbation of at most 16/255. Consistent with the original work [40], once the adversarial perturbation reaches a 90% confidence of fooling the detector, an Expectation over Transforms is introduced. This step applies random Gaussian blur and translation to the video before each adversarial step. The % AUC on adversarially-perturbed CDFv2, and visualization of a map of adversarial perturbations, is shown in Tab. III. The existing temporal DF detection methods rely on pixel intensity-based input without any constraints on the frequency components used for DF prediction. Consequently, this leads to a higher susceptibility to adversarial attacks – since typically such attacks reside in higher frequency bands [79]. In contrast, PhaseForensics shows better robustness since the band-pass phase is used (excluding the high-frequency components). Moreover, we observe that for adversarial robustness, alternate frequency decompositions, with more orientations and scales, show further improvements in adversarial robustness – as discussed next in Sec. IV-B.

B. Discussion and Ablation Studies

We now analyze our various design strategies related to: 1) configurations of CSP, 2) choice of face sub-region input, 3) choice of phase-based inputs (as compared to RGB-domain input, phase of discrete Fourier transform, amplitude of CSP), 4) choice of temporal preprocessing, and 5) choice of pre-training dataset. We also discuss the computational efficiency of our different models. We report the %AUC on CDFv2.

1) Computational Efficiency: Our default model, 4s2o-oct is a 36M-parameter model that has a video-friendly inference time (time for one forward pass) of 25.0ms on an NVIDIA Titan Xp GPU. As for the other versions of our model, with varying orientations and scales: 8s2o-half-oct is a 36.03M-parameter model with a inference time of 100.0ms for each forward pass, and 4s4-oct model is a 36.01M-parameter model that consumes 80.0ms for each forward pass. The increase in inference time is due to the increased number of input frequency sub-bands and the consequent computational overhead. Our default model is well-suited for video-based deepfake detection: our model can even be used to infer the authenticity of a live video stream in real-time. Notably, the inference time for our default model is comparable to that observed for some of the state-of-the-art methods. For example, FTCN has an average inference time of 27.5ms [85], and LipForensics is at 22.2ms [35]. All 3 methods can be used for real-time inference for a typical 30-fps video stream. On the other hand, based on the application (e.g., where one requires stronger robustness to attacks over speed or when real-time performance is not a necessity such as for offline videos), other versions of our model might be more suitable.

2) Design Choices for CSP: As stated in Sec. III, the CSP for lip region is computed with 4 scales, 2 orientations, and octave bandwidth filters We consider two additional CSP configurations: 1) 4 scales, 4 orientations, octave bandwidth filters, and 2) 8 scales, 2 orientations, half-octave filters. We evaluate adversarial robustness (Tab. III) and distortion robustness (Fig. 6) to color-based (contrast, and saturation) and compression-based (resizing and video compression) distortions with varying distortion levels. The %AUC for each plot in Fig. 6 is plotted as function of the distortion level for CDFv2. As is clear from Fig. 6, for distortion robustness, reducing the filter bandwidth (e.g., from octave to half-octave) does not show consistent performance gain for DF detector, since it increases the spatial support of the CSP filters [76], thereby reducing the fidelity to local motion. A similar observation holds for more than 2 orientations, since majority of lip motion tends to be captured by horizontal and vertical orientations. However, interestingly, increasing orientations and scales (by way of reducing filter bandwidth) improves adversarial robustness (as shown in Tab III), since these configurations allow for a more granular decomposition of the bandpass components.

3) PhaseForensics Applied to Face Regions: While lips provide a strong motion cue, in Tab. Va, we evaluate PhaseForensics applied to eyes and also full face, after pretraining on gaze prediction task (Sec. III) and lipreading respectively. As expected, learning a DF detector from eye regions using phase changes is difficult compared to lips, due to the nature of eye motion: more sporadic, and low amplitude compared to lips. Training with full face also shows a drop in performance since inputting face sub-regions (e.g. lips) allows the model to only look at relevant areas.

### TABLE IV

**Robustness to Multiple Attacks. Here We Show the Robustness of Various Temporal DF Detection Methods, Including PhaseForensics (Ours) to a Combination of Spatial Perturbations: One Color Distortion and Another Compression-Related**

| joint attack            | CNN-GRU | LipForensics | FTCN | ours |
|-------------------------|---------|--------------|------|------|
| saturation + resize     | 68.1    | 73.5         | 69.2 | **82.3** |
| contrast + resize       | 69.2    | 74.9         | 68.4 | **82.1** |
| saturation + compress   | 70.2    | 72.3         | 62.2 | **75.0** |
| contrast + compress     | 71.2    | 74.5         | 61.8 | **75.2** |
We motivate the design choice of our phase-based input in PhaseForensics by training DF detectors with alternate input feature choices, and evaluating on CDFv2. In this figure, we vary the CSP filter bandwidths, number of scales, and orientations for lip-region training, and observe the impact on varying levels of color and compression-related spatial distortions, starting with no distortion. Our experiments show that the most optimal configuration for the CSP is 4 scales, 2 orientations, and octave bandwidth.

![Figure 6](image_url)  
**Fig. 6.** We motivate the design choice of our phase-based input in PhaseForensics by training DF detectors with alternate input feature choices, and evaluating on CDFv2. In this figure, we vary the CSP filter bandwidths, number of scales, and orientations for lip-region training, and observe the impact on varying levels of color and compression-related spatial distortions, starting with no distortion. Our experiments show that the most optimal configuration for the CSP is 4 scales, 2 orientations, and octave bandwidth.

### TABLE V
**ADDITIONAL ANALYSIS.** (a) **WE SHOW THAT APPLYING PHASEFORENSICS TO LIP SUB-REGION YIELDS A BETTER PERFORMANCE COMPARED TO EYES OR FULL FACE.** (b) **WE VERIFY OUR CHOICE OF TEMPORAL PROCESSING OF PHASE USING SEPARABLE 3DCNN, IS THE MOST OPTIMAL CHOICE FOR TEMPORALLY PROCESSING PHASE-BASED INPUT, AS COMPARED TO NO TEMPORAL PROCESSING, AND TRADITIONAL 3DCNN.** WE **ALSO SHOW THAT SEPARABLE 3DCNN-BASED TEMPORAL PROCESSING IS THE MOST IMPACTFUL FOR PHASE-BASED FEATURES, AS COMPARED TO APPLYING SEPARABLE 3DCNN TEMPORAL PROCESSING TO RGB-DOMAIN INPUT. (c) **WE MOTIVATE THE CHOICE FOR PHASE FROM CSP BY COMPARING THE PERFORMANCE OF OTHERS KINDS OF INPUTS: PHASE FROM DISCRETE FOURIER TRANSFORM, AMPLITUDE OF CSP INCLUDING THE LOWPASS, BANDPASS AND HIGH-PASS COMPONENTS, AND THE ENTIRE CSP WITH PHASE FROM BANDPASS, AMPLITUDE OF ALL COMPONENTS. (d) **WE MOTIVATE THE CHOICE OF PRETRAINING ON LIPREADING, BY COMPARING AGAINST MODELS TRAINED ON KINETICS DATASET [8], AND NO PRETRAINING.**

| region | %AUC |
|--------|------|
| eye    | 73.3 |
| face   | 81.6 |
| lip    | 92.4 |

(Effect of face regions)

### TABLE V
**ADDITIONAL ANALYSIS.** (a) **WE SHOW THAT APPLYING PHASEFORENSICS TO LIP SUB-REGION YIELDS A BETTER PERFORMANCE COMPARED TO EYES OR FULL FACE.** (b) **WE VERIFY OUR CHOICE OF TEMPORAL PROCESSING OF PHASE USING SEPARABLE 3DCNN, IS THE MOST OPTIMAL CHOICE FOR TEMPORALLY PROCESSING PHASE-BASED INPUT, AS COMPARED TO NO TEMPORAL PROCESSING, AND TRADITIONAL 3DCNN.** WE **ALSO SHOW THAT SEPARABLE 3DCNN-BASED TEMPORAL PROCESSING IS THE MOST IMPACTFUL FOR PHASE-BASED FEATURES, AS COMPARED TO APPLYING SEPARABLE 3DCNN TEMPORAL PROCESSING TO RGB-DOMAIN INPUT. (c) **WE MOTIVATE THE CHOICE FOR PHASE FROM CSP BY COMPARING THE PERFORMANCE OF OTHERS KINDS OF INPUTS: PHASE FROM DISCRETE FOURIER TRANSFORM, AMPLITUDE OF CSP INCLUDING THE LOWPASS, BANDPASS AND HIGH-PASS COMPONENTS, AND THE ENTIRE CSP WITH PHASE FROM BANDPASS, AMPLITUDE OF ALL COMPONENTS.**

| input feature | %AUC |
|---------------|------|
| DFT phase     | 66.1 |
| CSP amplitude | 66.6 |
| Full CSP      | 88.4 |
| CSP phase     | 92.4 |

(Other input features)

| temporal proc. feature | %AUC |
|------------------------|------|
| 3DCNN, RGB [35]        | 62.4 |
| separable 3DCNN, RGB   | 83.1 |
| no processing, phase   | 62.8 |
| 3DCNN, phase           | 85.2 |
| separable 3DCNN, phase | 92.4 |

(Effect of temporal processing)

| pre-training | %AUC |
|--------------|------|
| no pretraining | 76.4 |
| Kinetics pretrain | 77.0 |
| upgrading pretrain | 92.4 |

(Choice of pretraining)

5) **Benefit of Phase in Comparison to Pixel-Intensity or Other Frequency-Domain Inputs:** Note that, the fifth row Tab. Vb (3DCNN) is a direct modification of the LipForensics pipeline: instead of training with pixel intensities as input (as done in LipForensics), we feed \( \Phi(t, x, y, c) \) from Eq. 2 to their model, appropriately adjusting the number of input channels. This allows for a clear assessment of the advantage of using phase instead of pixel values: compared to LipForensics, using phase inputs in the LipForensics pipeline improves the %AUC on CDFv2 from 82.4 to 85.2. As discussed above, further improvement is obtained by adopting Eq. 2 instead of the standard 3DCNN. We perform three additional analyses with three different kinds of inputs (shown in Tab. Vc): first, we replace the phase-based input from CSP with amplitude of CSP across all bandpass, low-pass, and high-pass components; second we experiment with the entire CSP, including phase and amplitude from all frequency sub-bands; and third, with the phase from the discrete Fourier transform. We observe sub-optimal results when only amplitude-based inputs of CSP are used, because these do not encode facial dynamics directly in the temporal variations as done by phase variations. Moreover, a CSP, given its finite support basis, isolates spatially-localized motion in the phase changes of frequency sub-bands – which is needed for local facial motion processing. However, the phase variations in a DFT capture only the global motion – making them less effective in local representing facial motion dynamics compared to the phase changes from a CSP.

6) **Performance of PhaseForensics on Lips, With Different Pretraining Strategies:** As detailed in Sec. III-B of the main paper, similar to LipForensics [35], pretraining on lipreading (using LRW dataset [16], Sec. III-B of main paper) enables us to learn a distribution of natural lip movements. Finetuning such a pretrained network allows us to effectively train a DF classifier by detecting deviations from natural lip movements. As shown in Tab. Vd, the performance of the PhaseForensics DF detector drops without the pretraining step. This is expected, since such a case is akin to expecting the DF detector to predict deviations from natural lip movements, without having seen the original distribution of natural lip movements first. In another experiments, we pretrain on Kinetics-400 action-recognition dataset as an alternative; this is also done.

4) **Temporal Preprocessing:** To motivate the importance of separable spatio-temporal phase preprocesssing (Sec. III), we train three models on the lip region: one without any spatio-temporal filtering (phase information is directly passed to ResNet-18 feature extractor), one with a standard 3DCNN layer (similar to LipForensics [35]), and one with our proposed separable 3D CNN layer. Looking at the performance on CDFv2, DFDC datasets in Tab. Vb, we conclude that our choice of separable spatio-temporal phase processing is most suited for PhaseForensics. Lastly, we would like to understand whether the biggest advantage of our method is in applying separable 3DCNN temporal processing or phase-based input. To assess this, we use a separable 3DCNN input layer for RGB-based inputs. The second row of Tab. Vb shows this experiment: we observe that while using separable 3DCNN temporal processing does lead to improved performance with RGB inputs, the performance improvement observed in the case of phase-based inputs is more apparent. This suggests that it is the combination of phase-based features as well as the effective temporal processing of these inputs using separable 3DCNN that results in the performance we observe for PhaseForensics.
in LipForensics [35]). We observe that this also leads to sub-optimal results, since the input domain is very different.

7) Limitations and Future Work: Similar to LipForensics [35], PhaseForensics is dependent on the presence of motion in the relevant facial regions (e.g., lips) and requires a domain-specific pretraining step. Moreover, CSP decompositions currently tend to form over-complete/redundant representations for images, which may create additional computational cost. More efficient alternatives such as Reisz pyramids [77] could be used in the future. Lastly, the distortion and adversarial robustness of PhaseForensics holds so long as these distortions do not disrupt the band-pasted frequency components or occlude facial regions: noise-like, blurring, and block-wise distortions sometimes do not satisfy this constraint. Moreover, this constraint is not only observed in our method, but also in state-of-the-art methods such as LipForensics [35] and FTCN [85] (see Tab. VI for average performance on noise, block-like, and blurring distortions applied to CDFv2 test set). However, such distortions show an obvious evidence of tampering: which already makes videos less reliable.  

V. CONCLUSION

In this work, we presented PhaseForensics, our method for generalizable, robust, deepfake detection. This method outperforms existing works with state-of-the-art cross-dataset generalization, and is robust under a variety of spatial and adversarial distortions. By being effective against both in and out of training domain samples, we take a step towards the real-world deployment of DF detectors.

ACKNOWLEDGMENT

The authors would like to thank Seanwook Park for help with setting up the EVE dataset, Alexandros Haliasos for discussion and clarifications on LipForensics, and Abhishek Badki and Orazio Gallo for the initial discussions.
Ekta Prashnani received the Ph.D. degree in electrical engineering from the University of California at Santa Barbara in 2022. She is currently a Research Scientist with NVIDIA Corporation, where her research focuses on media forensics algorithms, that aim to verify authorized use of synthetic avatars and foundation models for multi-modal human interactions. She has served as the Workshop Chair for the Inaugural Workshop on Computer Vision For Videogames at ECCV 2024 and on the program committee for prominent computer vision conferences and journals.

Michael Goebel received the Ph.D. degree in electrical engineering from the University of California at Santa Barbara in 2023. He is currently in the San Francisco Bay Area as a Software Engineer. His research interests include machine learning, computer vision, media forensics, and biomedical imaging.

B. S. Manjunath (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1991. He is currently a Distinguished Professor of electrical and computer engineering with the University of California at Santa Barbara, where he directs the Center for Multimodal Big Data Science and Healthcare. He has authored or co-authored over 350 peer-reviewed articles. His current research interests include image processing, machine learning, computer vision, and media forensics. He is a fellow of ACM, AIMBE, and the National Academy of Inventors (NAI). He served as an Associate Editor for several IEEE Transactions, including IEEE TRANSACTIONS ON IMAGE PROCESSING.