Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain

Honggu Liu[1]* Xiaodan Li[2] Wenbo Zhou[1]† Yuefeng Chen[2]
Yuan He[2] Hui Xue[2] Weiming Zhang[1]† Nenghai Yu[1]
[1]University of Science and Technology of China [2]Alibaba Group
lhg9754@mail.ustc.edu.cn, {welbeckz, zhangwm, ynh}@ustc.edu.cn
{fiona.lxd, yuefeng.chenyf, heyuan.hy, hui.xueh}@alibaba-inc.com

Abstract

*The remarkable success in face forgery techniques has received considerable attention in computer vision due to security concerns. We observe that up-sampling is a necessary step of most face forgery techniques, and cumulative up-sampling will result in obvious changes in the frequency domain, especially in the phase spectrum. According to the property of natural images, the phase spectrum preserves abundant frequency components that provide extra information and complement the loss of the amplitude spectrum. To this end, we present a novel **S**patial-**P**hase **S**hallow **L**earning (**SPSL**) method, which combines spatial image and phase spectrum to capture the up-sampling artifacts of face forgery to improve the transferability, for face forgery detection. And we also theoretically analyze the validity of utilizing the phase spectrum. Moreover, we notice that local texture information is more crucial than high-level semantic information for the face forgery detection task. So we reduce the receptive fields by shallowing the network to suppress high-level features and focus on the local region. Extensive experiments show that SPSL can achieve the state-of-the-art performance on cross-datasets evaluation as well as multi-class classification and obtain comparable results on single dataset evaluation.*

1. Introduction

Benefiting from the tremendous success of generative techniques, such as Variational Autoencoders (VAE) [34] and Generative Adversarial Networks (GANs) [15], face forgery has become an emerging hot research topic in very recent years. The face forgery techniques are able to synthesize realistic faces that are indistinguishable for human eyes. However, these forgery techniques are likely to be abused for malicious purposes, causing serious security and ethical issues (*e.g.* celebrity pornography and political persecution). Therefore, it is of paramount importance to develop more general and practical methods for face forgery detection.

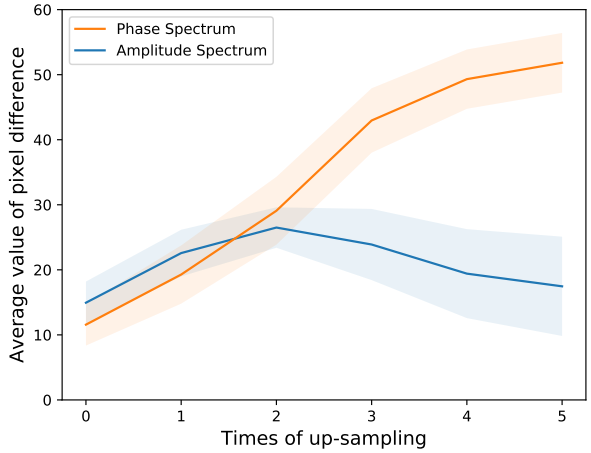


Figure 1: **The variation analysis in the frequency domain**. The curve shows the average value of pixel difference (mean and variance) of the phase spectrum between 1000 origin and up-sampling samples increase dramatically with the increase of the number of up-sampling.

To alleviate the risks brought by malicious usage of face forgery, various methods [49, 25, 1, 48, 2, 38, 37, 12, 22, 35, 28, 8, 23] have been proposed. Most of these methods detect face forgery in a supervised fashion with prior knowledge of face manipulation methods [1, 6, 23]. Under this setting, these approaches achieve excellent performance on some public datasets [48, 20, 37, 26, 45]. However, these detection methods tend to suffer from overfitting thus their effectiveness is limited to the datasets which they are specifically trained on. Moreover, in real-world detection, it is inevitable to face a source/target mismatch

* Work done during the internship of Honggu Liu at Alibaba Group.
† Corresponding Author.

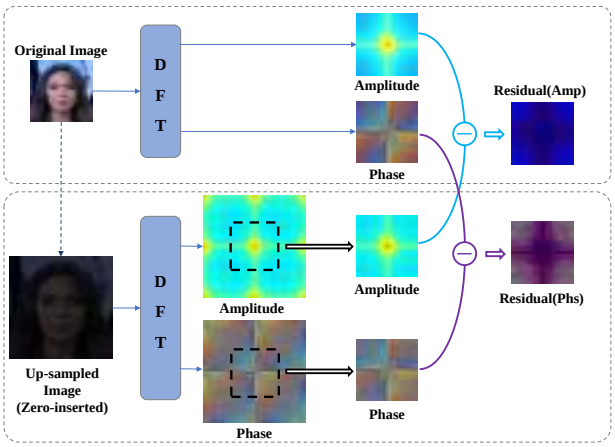


Figure 2: The frequency domain analysis of the origin and up-sampling face. The residual images show that the differences of phase spectrum between origin and up-sampling are bigger than the amplitude spectrum with the up-sampling. Best viewed in color. (Darker color indicates smaller pixel values).

problem when handling unseen samples, which is extremely challenging for deepfake detection task. Therefore, it is necessary to enhance the generalization and transferability of forgery detection techniques. Recently, some approaches [10, 22, 28, 46] have made attempts to improve the transferability, there are still deficiencies. For example, two-branch [28] achieves the state-of-the-art performance of cross-dataset detection at the expense of frame-level detection accuracy. Face X-ray [22] starts from a novel perspective that aims at detecting the blending boundary artifacts and obtains perfect performances in detecting unseen forgery method for raw videos, which significantly improves transferability of different manipulation methods [9, 43, 13, 42]. However, since only focus on the information extracted from the spatial domain, it is easily to be influenced by video compression. Thus, we need to concern about the more common artifacts in the generation of forgery images from various domains.

We obverse that **up-sampling** is a non-negligible step in generative models(*e.g.* VAE [34], GANs [15]), from which the generated part are then used to synthesize fake faces. This operation usually leaves a trace in the frequency domain, which provides cues for separating synthesized faces from real ones. Though [11] also tried to detect these artifacts with the amplitude spectrum, the performance is limited due to the information loss.

To this end, we propose Spatial-Phase Shallow Learning (SPSL) for face forgery detection, which leverages the phase spectrum for detecting the common artifacts. The pivotal thought is that the phase spectrum is more hypersensitive to up-sampling than the amplitude spectrum. As shown in Figure 1, with more times of up-sampling operations are performed, the average pixel differences of the phase spectrum get much greater than that of the amplitude spectrum. A visualization comparison can also be found in Figure 2.

Moreover, [5] used the patch-based classification to generalize the image forensics. Thus, we hold the opinion that local texture information is more important than high-level semantic information even high-level semantic information should be suppressed to a certain extent in the specific task of forged face detection. For this purpose, SPSL drops many convolutional layers to reduce the receptive field [3] and forces CNNs to pay more attention to local regions which are abundant in textures and lack high-level semantic information.

As a result, SPSL focuses on the common step in the forged faces generation and pays more attention to textures leading to a performance improvement of the cross-datasets evaluation. At the same time, the performance on multi-class classification also improves due to the specific traces of different categories manipulation are left in the phase spectrum.

Extensive experiments demonstrate that SPSL significantly improved the transferability and achieved the state-of-the-art over cross-dataset evaluation.

The major contributions in this paper are summarized as follows:

- We firstly leverage the phase spectrum to detect forged face images and demonstrate that CNNs can capture extra implicit features of the phase spectrum which are beneficial to face forgery detection with precise mathematical derivation.
- Aiming at the specific problem of forged face detection, we assume that high-level semantic information should be appropriately suppressed. And we experimentally validate the hypothesis by decreasing the receptive field of CNNs with shallow network learning.
- We verify that our approach achieves the state-of-the-art performance of forged face detection over cross-dataset evaluation.

2. Related work

Since face manipulation is a classical research topic [43, 33, 32, 21, 17, 42] in computer vision, verifying its authenticity is not a new problem. However, recent remarkable successes of deep learning make face manipulation easier and more realistic which poses a significant challenge of forged face detection. In this section, we briefly review the current face forgery detection methods that are representative and related to our work.

2.1. Spatial-based Face forgery Detection

With the development of face forgery, a wide variety of methods have been proposed to detect forged face. The majority of them exploit artifacts based on the spatial domain, especially in RGB. Some methods for deepfake detection focus on hand-crafted facial features from the video, such as eye blinking [24], inconsistent head poses [48], facial expression change [2]. Recent methods [1, 31, 37] capture high-level features from the spatial domain by using deep neural networks and show impressive performance. Nguyen *et al*. proposed a method [31] which leveraging capsule network [39] to detect face manipulation. Rossler *et al*. [37] show the best performance on many kinds of forgery algorithms with the efficient XceptionNet [6] at that time. Face X-ray [22] mainly focuses on the blending step which exists in most face forgery and thus achieved state-of-the-art performance on transferability in raw videos. However, it still has some limitations that the performance of Face X-ray will sharply drop when encounter low-resolution images, and it may not work with entirely synthesized images. Almost all of these CNN-based methods only use spatial domain information and therefore the performance is quite sensitive to the quality or data distribution of datasets. In our work, we combine the spatial domain with the frequency domain to take advantage of both.

2.2. Frequency-based Face forgery Detection

Besides focusing on the spatial domain, some methods pay attention to the frequency domain for capturing artifacts of the forgery. In fact, frequency analysis is a common and important way in digital image processing and has been widely applied to various tasks in computer vision [44, 47, 41, 18]. Most of them use either Discrete Fourier Transform (DFT) or Wavelet Transform (WT), or Discrete Cosine Transform (DCT) to convert the spatial image to the frequency domain. Durall *et al*. [12] first proposed that averaging the amplitude of each frequency band with DFT can mine abnormal information of forgery in face manipulation detection. F^3-Net [35] extracted frequency-domain information using DCT and analyzed the statistic features for face forgery detection. F^3-Net achieved state-of-the-art performance on highly compressed videos, but the performance on cross-dataset evaluation drops greatly. Masi *et al*. [28] leverage a Laplacian of Gaussian (LoG) to make frequency enhancement for purpose of suppressing the image content present in the low-level feature maps. However, most of these related works mainly depend on low-level statistical features rather than high-level features extracted by CNNs, and therefore the frequency information was inadequately utilized. In our work, given the powerful capabilities of feature extraction of CNNs, we make use of the phase spectrum in DFT and explicitly prove the validity with theory analysis while we integrate it into the whole learning process of CNNs.

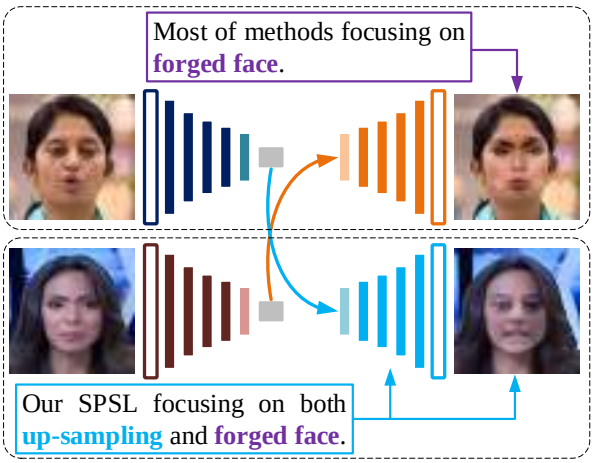


Figure 3: Overview of typical face manipulation pipeline. Most of the previous works only focus on forged faces, while we focus on both up-sampling and forged faces.

3. SPSL for Face Forgery Detection

In this section, we start by introducing the key observation of face forgery generation. Then we propose spatial-phase shallow learning to detect the observed common artifacts for face forgery detection. Finally, Making the network shallow can focus on the local region to further boost the improvement of transferability.

3.1. Motivation

As shown in Figure 3, a typical facial manipulation method consists of three stages [9]: 1) encoding source face; 2) swapping face in latent space; 3) decoding target face.

Up-sampling is a vital step for decoding the target face, based on either AutoEncoders [34] or GANs [15]. Thus, we leverage the phase information to detect up-sampling artifacts. For applying phase information to CNNs, we reconstruct the spatial domain representation of the phase spectrum from the frequency domain (*i.e*. IDFT with the frequency spectrum without amplitude). Finally, we concatenate the spatial domain representation of the phase spectrum with the RGB image in the channel, which results in an RGBP 4-channel image.

3.2. Capturing up-sampling artifacts via phase spectrum in face forgery

To detect the observed common artifacts, namely up-sampling, we analyze it in the frequency domain.

Up-sampling will lead to the emergence of new frequency components. And we make a claim as follows

Claim 1. *Phase spectrum is more sensitive to up-sampling artifacts and therefore helps face forgery detection.*

To simplify the calculation, we make all the mathematical derivation based on the 1D signal. We first set up the basic notations used in this paper: $x(n)$ and $\mathbf{X}(u)$ denote a 1D discrete signal and its Discrete Fourier Transform(DFT), where n is the spatial location of the signal and u represents the frequency. $\mathbf{A}(u)$ is the amplitude spectrum and $\mathbf{P}(u)$ is the phase spectrum. And we use $c(n)$ and $\mathbf{C}(n)$ to denote the convolution kernel and its DFT representation. And we use $*$ to denote convolutional operation.

Proof. The increase of spatial resolution in 2D corresponds to the extension of the time domain in 1D. Assume that the input $x(n)$ is up-sampled by factor 2, then

$$\hat{x}(n) = \begin{cases} x(\frac{1}{2}n), & n = 2k \\ 0, & n = 2k+1 \end{cases} \tag{1}$$

where $k = 0, 1, 2, \cdots, N-1$, and

$$\begin{aligned} \hat{\mathbf{X}}(u) &= \frac{1}{2N}\sum_{n=0}^{2N-1}\hat{x}(n)e^{-j\frac{2\pi un}{2N}} \\ &= \frac{1}{2N}\sum_{n=0}^{N-1}\hat{x}(2n)e^{-j\frac{2\pi u2n}{2N}} \\ &= \frac{1}{2N}\sum_{n=0}^{N-1}x(n)e^{-j\frac{2\pi 2un}{N}} \end{aligned} \tag{2}$$

The we have $\hat{x}(n) = x(\frac{1}{2}n) \Leftrightarrow \hat{\mathbf{X}}(u) = \mathbf{X}(2u)$ with the Eq. 2, which leads to the conclusion that the increase of spatial resolution will result in the compression in the frequency domain which is consistent with the property of Fourier Transform (FT). In fact, the essence of DFT is the principle value interval of Discrete Fourier Series (DFS) and thus new frequency components are the duplicate of origin frequency components.

Base on our inference that phase spectrum will keep more frequency components that tend to zero in amplitude spectrum, which is detailedly proved in the **Appendix** 1.1. We first assume the amplitude spectrum $\mathbf{X_A}(u)$ and the phase spectrum $\mathbf{X_P}(u)$ of original images $x(n)$. It is

$$\begin{aligned} \mathbf{X}_A(u) &= \underbrace{a_0 + a_1e^{j\theta_1} + \cdots + a_ke^{j\theta_k}}_{(k+1)\text{ items}} \\ \mathbf{X}_P(u) &= \underbrace{p_0 + p_1e^{j\theta_1} + \cdots + p_{N-1}e^{j\theta_{N-1}}}_{N\text{ items}} \end{aligned} \tag{3}$$

and the corresponding up-sampling is

$$\begin{aligned} \mathbf{X}_A^{up}(u) &= \underbrace{a_0 + \cdots + a_ke^{j\theta_k}}_{(k+1)\text{ items}} + \underbrace{a_Ne^{j\theta_N} + \cdots + a_{N+k}e^{j\theta_{N+k}}}_{(k+1)\text{ items}} \\ \mathbf{X}_P^{up}(u) &= \underbrace{p_0 + p_1e^{j\theta_1} + \cdots + p_{2N-1}e^{j\theta_{2N-1}}}_{2N\text{ items}} \end{aligned} \tag{4}$$

We define that $y_A(n)$ is the output of a convolution layer with an input $x(n)$ and its frequency domain form is $\mathbf{Y_A}(u)$. And we get

$$\begin{aligned} y_A(n) &= x(n) * c(n) \\ &\Updownarrow \\ \mathbf{Y}_A(u) &= \mathbf{X_A}(u) \cdot \mathbf{C}(u) \end{aligned} \tag{5}$$

According to the deduction that the phase spectrum helps CNNs acquire and learn more abundant frequency components which are ignored with convolution calculations of amplitude spectrum proved in **Appendix** 1.2, we can deduce the frequency domain form is

$$\mathbf{Y}_A(u) = \underbrace{f_0 + f_1e^{j\theta_1} + \cdots + f_{k+N-1}e^{j\theta_{k+N-1}}}_{k+N\text{ items}} \tag{6}$$

$$\mathbf{Y}_A^{up}(u) = \underbrace{f_0 + f_1e^{j\theta_1} + \cdots + f_{2N+k-1}e^{j\theta_{2N+k-1}}}_{2N\text{ items}} \tag{7}$$

In our work, we first take Inverse Discrete Fourier Transform (IDFT) to phase spectrum and acquire the spatial domain form $p(n)$ of phase. And we state a theorem named *the distributive law* as follow,

Theorem 1. $(f(\cdot) + g(\cdot)) * h(\cdot) = f(\cdot) * h(\cdot) + g(\cdot) * h(\cdot)$

Then we consider that we directly concatenate $x(n)$ and $p(n)$ in channel dimension based on theorem 1 and the output $\mathbf{Y}_{A+P}(u)$ is

$$\mathbf{Y}_{A+P}(u) = \underbrace{f_0 + f_1e^{j\theta_1} + \cdots + f_{2N-2}e^{j\theta_{2N-2}}}_{2N-1\text{ items}} \tag{8}$$

$$\mathbf{Y}_{A+P}^{up}(u) = \underbrace{f_0 + f_1e^{j\theta_1} + \cdots + f_{3N-2}e^{j\theta_{3N-2}}}_{3N-1\text{ items}} \tag{9}$$

■

Intuitively, the number of learnable frequency components is N when we leverage the original image and its phase together, but the number just is $N-k$ with the utilization of the original image alone. Therefore, it is clear that the difference between $\mathbf{Y}_{A+P}(u)$ and $\mathbf{Y}_{A+P}^{up}(u)$ is bigger than $\mathbf{Y}_A(u)$ and $\mathbf{Y}_A^{up}(u)$. In particular, the value of k is usually small in nature images and thus our method observably improves the performance on the detection of face forgery. Thus, we conclude that the introduction of the phase spectrum helps capture more frequency artifacts caused by cumulative up-sampling in deepfake video generation.

3.3. Suppressing the semantic information and focusing on local region

We consider that the pivotal distinction between pristine face and forged face is local low-level features(*e.g.* textures, colors) instead of global high-level semantic features(*e.g.*

Methods	HQ		LQ	
	ACC	AUC	ACC	AUC
Steg. Features [14]	70.97	-	55.98	-
Cozzolino *et al*. [7]	78.45	-	58.69	-
Bayer & Stamm [4]	82.97	-	66.84	-
Rahmouni *et al*. [36]	79.08	-	61.18	-
MesoNet [1]	83.10	-	70.47	-
Face X-ray [22]	-	87.35	-	61.60
Xception [6]	**92.39**	94.86	80.32	81.76
Ours(Xception)	91.50	**95.32**	**81.57**	**82.82**

Table 1: Quantitative results (ACC (%) and AUC (%)) on FaceForensics++ dataset with high-quality (light compression) and low quality (heavy compression) settings. The bold results are the best.

face, human). Because most of these semantic features are shared in both pristine and forged faces, extensive high-level semantic information more or less has a negative effect on forged face detection as it contains many common characteristics of pristine and forged face images. For the purpose of suppressing high-level semantic features and extracting more texture features, we straightforwardly shallow the neural network by throwing away many convolutional layers or blocks. Then we demonstrate that shallow networks are more transferable and efficient simultaneously.

4. Experiments

In this section, we first introduce the overall experimental settings and then present extensive experimental results to demonstrate the superiority of our approach.

4.1. Experimental settings

Datasets. Following recent related works [35, 28, 22, 23] of face forgery detection, we conduct our experiments on the two benchmark public deepfake datasets: FaceForensics++(FF++) [37] and Celeb-DF [26]. Both of them are large-scale and contain pristine and manipulated videos of human faces. FF++ consists of four kinds of common face manipulation methods [9, 43, 13, 42]. Celeb-DF is in general the most challenging to the current detection methods, and their overall performance on Celeb-DF is lowest across all datasets.

Evaluation metrics. In our experiments, we mainly utilize the Accuracy rate (ACC) and the Area Under Receiver Operating Characteristic Curve (AUC) as our evaluation metrics. **(1) ACC.** Accuracy rate is the most intuitive metric in face forgery detection. It is also applied to FF++ [37] and thus we use ACC as the major evaluation metric in the experiment. **(2) AUC.** Following the Celeb-DF [26] and Two-branch [28], we use AUC as another evaluation metric to evaluate the performance on cross-dataset. Besides, we use the recall rate as our multi-class classification evaluation metric.

Implementation and Hyper-Parameters. In our experiments, we use Xception [6] as the backbone of our approach. For the purpose of reducing receptive fields, we just retain the Xception Block 1-3 and Xception Block 12. The final spatial form of our phase spectrum is the IDFT of the absolute value of the pristine phase spectrum. We optimize the networks by Adam optimizer [19]. The initial learning rate $lr = 2 \times 10^{-3}$ and it drops to half of itself every time the validation loss does not decrease after 5 full epochs.

4.2. Comparison with previous methods

In this section, we compare our method with previous deepfake detection methods. We train all models on only FF++ [37] and respectively evaluate them on FF++ in Section 4.2.1 and Celeb-DF in Section 4.2.2.

4.2.1 Comparable results on FF++

In this section, we compare our method with previous deepfake detection methods on FF++ [37]. Although our primary purpose is to improve the generalization and transferability, we also obtain comparable results in FF++ [37]. We first evaluate our methods on different video compression settings including high quality (HQ (c23)) and low quality (LQ (c40)). As the results shown in Table 1, the proposed method outperforms or is on par with baseline in both ACC and AUC with LQ settings. Low-quality videos have been highly compressed and many frequency components are weakened. The improvement of performance mainly benefits from the extra phase information captured by CNNs, which keeps more frequency components than plain RGB-based images. At the same time, we also obtain comparable results with HQ settings.

Furthermore, we also evaluate our approach on different face manipulation methods in FF++ [37]. The results are demonstrated in Table 2. We train and test our models exactly on low-quality videos for each manipulation methods. We also reproduced the results of MesoNet [1] and Xception [6], and other results are directly cited from [37]. In general, basic experiments also show comparable results with previous methods though the transferability is the main purpose of SPSL.

4.2.2 Cross-dataset evaluation on Celeb-DF

In this section, we evaluate the transferability of our method given that it is trained on FF++ with multiple manipulations but tested on Celeb-DF. We first verify that most of the previous methods show a drastic performance drop on the cross-dataset evaluation. Table 3 shows the AUC comparison with some recent methods for face forgery detection. Our method obtains the state-of-the-art AUC on Celeb-DF

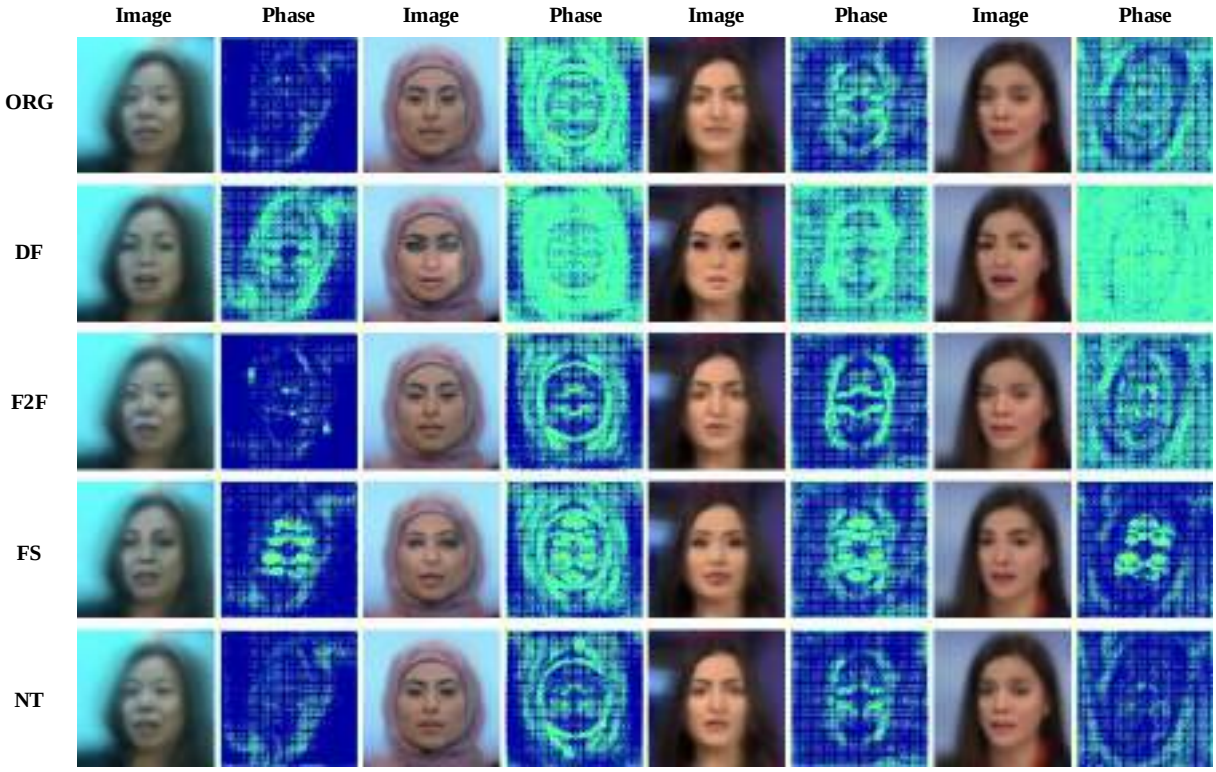


Figure 4: Visualization of various manipulation methods and our spatial domain representations of the phase spectrum in FF++ [37]. Each image and phase spectrum is the average of all frames of a video. Every manipulation method tends to be a specific pattern in the phase spectrum while it is not obvious in the RGB domain. Best viewed in color.

Methods	DF [9]		F2F [43]		FS [13]		NT [42]	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Steg. Features [14]	73.64	-	73.72	-	68.93	-	63.33	-
Cozzolino *et al*. [7]	85.45	-	67.88	-	73.79	-	78.00	-
Rahmouni *et al*. [36]	85.45	-	64.23	-	56.31	-	60.07	-
Bayar and Stamm [4]	84.55	-	73.72	-	82.52	-	70.67	-
MesoNet [1]	87.27	-	56.20	-	61.17	-	40.67	-
XceptionNet [6]	**95.15**	**99.08**	83.48	93.77	92.09	97.42	**77.89**	**84.23**
Ours(Xception)	93.48	98.50	**86.02**	**94.62**	**92.26**	**98.10**	76.78	80.49

Table 2: Quantitative results (ACC (%) and AUC (%)) on FaceForensics++ dataset with four different manipulation methods, *i.e*. DeepFakes(DF) [9], Face2Face(F2F) [43], FaceSwap(FS) [13], NeuralTextures(NT) [42]. The bold results are best.

while still having a good performance on FF++ compared with all the other methods. The performance gains mainly benefit from the extra appreciable frequency components of the phase spectrum, which are enhanced many times in the cumulative up-sampling, and thus the differences of frequency components perceptible by convolutional kernel between pristine images and forgery become more striking. Therefore, the proposed SPSL is capable of detecting common artifacts.

4.3. Multi-class classification evaluation

We furthermore evaluate the proposed SPSL on multi-class classification with different face manipulation methods list in FF++ [37] in this section. The models are trained and tested on FF++ with five types of labels, and multi-class classification is more challenging and significant than binary classification. The results are shown in Table 4 by the way of recall rate. With all three kinds of compression setting, the proposed SPSL completely surpasses the orig-

Method	FF++ [37]	Celeb-DF [26]
Two-stream [49]	70.10	53.80
Meso4 [1]	84.70	54.80
MesoInception4 [1]	83.00	53.60
HeadPose [48]	47.30	54.60
FWA [25]	80.10	56.90
VA-MLP [29]	66.40	55.00
VA-LogReg	78.00	55.10
Xception-c40 [37]	95.50	65.50
Multi-task [30]	76.30	54.30
Capsule [31]	96.60	57.50
DSP-FWA [25]	93.00	64.60
SMIL [23]	96.80	56.30
Two-branch [28]	93.20	73.40
F^3-Net [35]	**97.97**	65.17
SPSL(Xception)	96.91	**76.88**

Table 3: **Cross-dataset evaluation (AUC (%)) on Celeb-DF**. Best competing methods on Celeb-DF are reported. Our method obtains the state-of-the-art performance on cross-dataset evaluation. At the same time, our method still performs well when tested on just deepfake class(96.91%) AUC on FF++. Results for some other methods are from [26], and the bold results are the best.

inal XceptionNet method. Furthermore, we also show the t-SNE [27] feature spaces of data in FF++ high quality with the multi-class classification task, by the Xception and our SPSL, as shown in Figure 5. Xception is more likely to confuse pristine faces with NeuralTextures-based fake faces because this manipulation method modifies very limited pixels in the spatial domain, as shown in Figure 5(a). Conversely, the proposed SPSL can split up all classes in the embedding feature spaces, as shown in Figure 5(b). These improvements may benefit from the salient difference of phase spectrum among various manipulation methods and we show the average phase spectrum in the spatial domain of every frame of a video in Figure 4. For all four kinds of manipulation methods in FF++ [37], the spatial images of the phase spectrum show distinguishing results for each method. In particular, NeuralTextures-based images, which just slightly tampered with lip, are very similar to pristine images causing almost indistinguishable in the RGB domain but the spatial images of the phase spectrum are still separable.

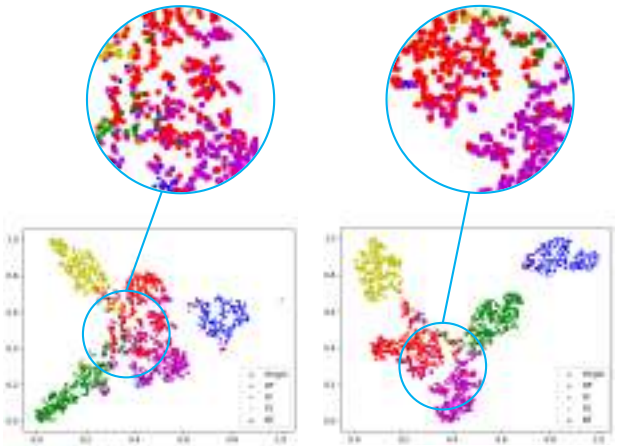
5. Ablation Study

5.1. Effectiveness of Phase spectrum and Shallow network

To evaluate the effectiveness of both Phase spectrum and Shallow network, we first respectively evaluate one of them with baseline and combine them finally. All models are trained on FF++ [37] and tested on Celeb-DF [26]. The results are listed in Table 5. Compared with model 1(baseline Xception), model 2(Xception with phase spectrum) and model 3 (shallow Xception) improve the AUC scores of Celeb-DF. The transferability has a great improvement with both of them. When combining phase spectrum and shallow network, SPSL gets the best performance and the AUC score increase by about 13%. Furthermore, to demonstrate the effectiveness of the proposed SPSL better, we respectively visualize the Gradient-weighted Class Activation Mapping(Grad-CAM) [40] of the baseline and SPSL,

Methods	DF	F2F	FS	NT	ORG
MesoIncep4 [1]	94.81	43.32	73.24	40.39	85.16
Xception-c0 [6]	97.84	96.68	96.84	87.67	98.03
SPSL (Xception-c0)	**99.05**	**97.20**	**97.63**	**91.40**	**98.25**
Xception-c23 [6]	88.00	88.61	87.07	74.83	75.52
SPSL (Xception-c23)	**94.18**	**93.59**	**95.62**	**81.72**	**88.72**
Xception-c40 [6]	86.61	**78.88**	83.16	52.94	75.55
SPSL (Xception-c40)	**91.16**	78.31	**88.75**	**58.97**	**77.49**

Table 4: The recall rate (%) of origin and each manipulation method with Raw (c0) , HQ (c23) and LQ (c40) settings in our multi-class classification.



(a) Baseline (b) SPSL

Figure 5: The t-SNE feature spaces visualization of the basic Xception (a) and SPSL (b) on FaceForensics++ [37] high quality (HQ) in the multi-class classification task. Red color dots represent pristine images, and rest colors respectively indicate the different manipulation methods. Best viewed in color.

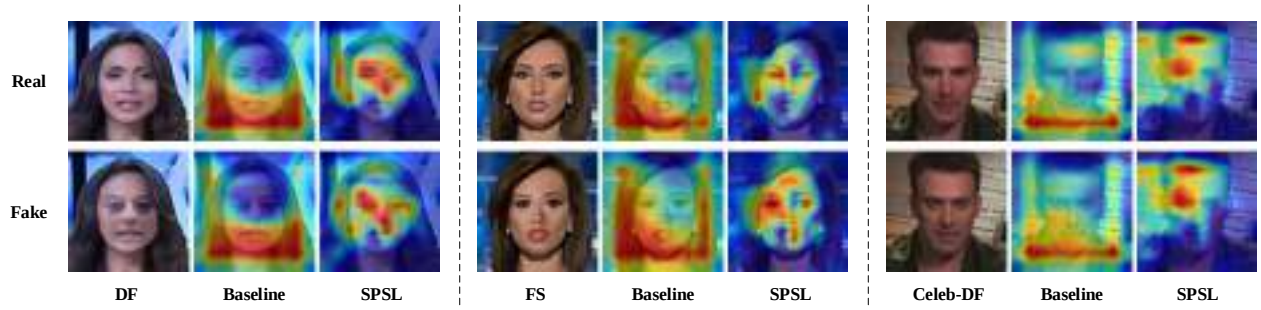


Figure 6: The Grad-CAM of the baseline Xception and the proposed SPSL, including two different manipulation methods in FF++ [37] and another Celeb-DF [26] datasets. Best viewed in color.

as shown in Figure 6. The Grad-CAM indicates that the proposed SPSL prefers to focus on more microcosmic regions while the baseline model pays more attention to global information, and this phenomenon also accords with our motivation. Furthermore, we demonstrate the correlation analysis between the performance and the number of convolution layers of various backbone networks in **Appendix** 2.

ID	Phase	Shallow	Celeb-DF [26]
1	-	-	59.98
2	√	-	69.01
3	-	√	66.74
4	√	√	**72.39**

Table 5: Ablation study of the proposed SPSL. These models are trained on FF++ with high quality(HQ) settings and tested on Celeb-DF (AUC (%)). We compare SPSL and its variants by removing phase spectrum and shallow operation step by step.

5.2. Universality of SPSL with Various Backbones

All the above-mentioned experiments are based on XceptionNet [6], and thus we also evaluate the universality of SPSL with two types of ResNet [16]. For both ResNet34 and ResNet50, we directly halve the residual block to shallow networks. The results listed in Table 6 demonstrate that the proposed SPSL is a general framework for various backbones.

6. Limitations

Even if we have demonstrated the effectiveness of the proposed SPSL and achieved satisfactory performance on cross-dataset evaluation and multi-class classification, we are aware that there exist some limitations of our work.

Our method depends on the existence of up-sampling in forged face generation. Thus, the performance may drop if the forgery face is not produced by methods based on generative models. Besides, our method also suffers from a transferability drop when encountering an entirely different type of face forgery manipulation. For instance, the model trained on identity swap datasets may fail to detect forgery faces whose expressions are swapped. This is expected since manipulations from different categories can leave a specific trace in the phase spectrum as shown in Figure 4, this is also the reason why our method makes a remarkable improvement of multi-class classification.

Backbone	FF++ [37]		Celeb-DF [26]	
	ACC	AUC	ACC	AUC
ResNet-34	71.55	81.58	65.19	66.90
SPSL (ResNet-34)	**83.24**	**89.26**	**66.79**	**71.78**
ResNet-50	81.83	83.51	**69.40**	70.05
SPSL (ResNet-50)	**86.64**	**91.04**	68.28	**73.09**

Table 6: The results (ACC (%) and AUC (%)) on FF++ [37] and cross-dataset evaluation on Celeb-DF [26] of two different backbones with the proposed SPSL.

7. Conclusion

In this work, we propose a novel face forgery detection method, SPSL, which takes advantage of both spatial and frequency information. The core competence of SPSL is that phase spectrum contains more abundant appreciable frequency components and these components will be duplicated in the process of up-sampling which is the necessary step of forged face generation. Besides, SPSL forces the network to focus on the local microcosmic region and suppress global semantic information for more robustness. We perform a meticulous mathematical derivation to prove the rationality of the proposed SPSL, and extensive experiments demonstrate that the SPSL has an excellent performance on the face forgery detection, especially in the challenging cross-dataset evaluation task.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 1, 3, 5, 6, 7

[2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, pages 38–45, 2019. 1, 3

[3] André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 2019. https://distill.pub/2019/computing-receptive-fields. 2

[4] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016. 5, 6

[5] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. *arXiv preprint arXiv:2008.10588*, 2020. 2

[6] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3, 5, 6, 7, 8

[7] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pages 159–164, 2017. 5, 6

[8] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020. 1

[9] DeepFakes. Deepfakes github. http://github.com/deepfakes/faceswap, 2017. Accessed 2020-08-18. 2, 3, 5, 6

[10] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards generalizable forgery detection with locality-aware autoencoder. *arXiv preprint arXiv:1909.05999*, 2019. 2

[11] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7890–7899, 2020. 2

[12] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019. 1, 3

[13] FaceSwap. Faceswap github. https://github.com/MarekKowalski/FaceSwap, 2016. Accessed 2020-08-18. 2, 5, 6

[14] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 5, 6

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2, 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[18] Gahyun Kim, Sungmin Eum, Jae Kyu Suhr, Dong Ik Kim, Kang Ryoung Park, and Jaihie Kim. Face liveness detection based on texture and frequency analyses. In *2012 5th IAPR international conference on biometrics (ICB)*, pages 67–72. IEEE, 2012. 3

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 1

[21] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017. 2

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 1, 2, 3, 5

[23] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1864–1872, 2020. 1, 5, 7

[24] 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. IEEE, 2018. 3

[25] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 1, 7

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 1, 5, 7, 8

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[28] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. *arXiv preprint arXiv:2008.03412*, 2020. 1, 2, 3, 5, 7

[29] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manip-

ulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019. 7

[30] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. 7

[31] Huy H Nguyen, Junichi Yamagishi, and Isao 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. IEEE, 2019. 3, 7

[32] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE international conference on computer vision*, pages 7184–7193, 2019. 2

[33] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Jian Jiang, Luis RP, Sheng Zhang, Pingyu Wu, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. 2

[34] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016. 1, 2, 3

[35] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 86–103, Cham, 2020. Springer International Publishing. 1, 3, 5, 7

[36] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017. 5, 6

[37] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 1, 3, 5, 6, 7, 8

[38] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1), 2019. 1

[39] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017. 3

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[41] José A Stuchi, Marcus A Angeloni, Rodrigo F Pereira, Levy Boccato, Guilherme Folego, Paulo VS Prado, and Romis RF Attux. Improving image classification with frequency domain layers for feature extraction. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017. 3

[42] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 5, 6

[43] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2, 5, 6

[44] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 3

[45] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 1

[46] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the generalization of gan image forensics. In *Chinese Conference on Biometric Recognition*, pages 134–141. Springer, 2019. 2

[47] Shengke Xue, Wenyuan Qiu, Fan Liu, and Xinyu Jin. Faster image super-resolution by improved frequency-domain neural networks. *Signal, Image and Video Processing*, 14(2):257–265, 2020. 3

[48] Xin Yang, Yuezun Li, and Siwei 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. IEEE, 2019. 1, 3, 7

[49] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry 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. IEEE, 2017. 1, 7