Beyond the Spectrum: Detecting Deepfakes via Re-Synthesis

Yang He[1], **Ning Yu**[2,3], **Margret Keuper**[4], **Mario Fritz**[1]
[1]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[2]Max Planck Institute for Informatics, Saarbrücken, Germany
[3]University of Maryland, College Park, United States
[4]University of Mannheim, Mannheim, Germany
yang.he@cispa.saarland, ningyu@mpi-inf.mpg.de, keuper@uni-mannheim.de, fritz@cispa.saarland

Abstract

The rapid advances in deep generative models over the past years have led to highly realistic media, known as deepfakes, that are commonly indistinguishable from real to human eyes. These advances make assessing the authenticity of visual data increasingly difficult and pose a misinformation threat to the trustworthiness of visual content in general. Although recent work has shown strong detection accuracy of such deepfakes, the success largely relies on identifying frequency artifacts in the generated images, which will not yield a sustainable detection approach as generative models continue evolving and closing the gap to real images. In order to overcome this issue, we propose a novel fake detection that is designed to re-synthesize testing images and extract visual cues for detection. The re-synthesis procedure is flexible, allowing us to incorporate a series of visual tasks - we adopt super-resolution, denoising and colorization as the re-synthesis. We demonstrate the improved effectiveness, cross-GAN generalization, and robustness against perturbations of our approach in a variety of detection scenarios involving multiple generators over CelebA-HQ, FFHQ, and LSUN datasets. Source code is available at https://github.com/SSAW14/BeyondtheSpectrum.

1 Introduction

In the past years, image generation and tampering techniques have been evolving quickly, benefiting from the continuous breakthroughs in generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] and its variations [Radford *et al.*, 2016; Gulrajani *et al.*, 2017; Karras *et al.*, 2018; Karras *et al.*, 2019; Karras *et al.*, 2020; Yu *et al.*, 2021a]. The fidelity and diversity of generated images have improved to a level that is arguably already photorealistic. Although fostering the techniques for numerous novel applications [Reed *et al.*, 2016; Thies *et al.*, 2016; Zhu *et al.*, 2017; Choi *et al.*, 2018; Yu *et al.*, 2019a; Yu *et al.*, 2020; Wang *et al.*, 2021], this development, on the other hand, poses new risks as recent results are challenging to be distinguished from real images by human eyes.

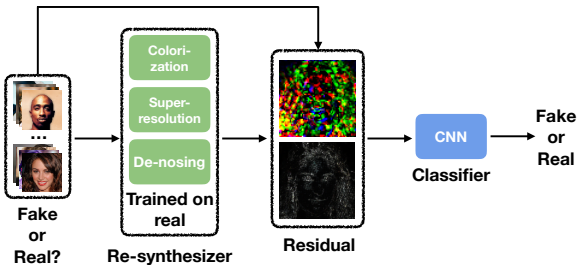

Figure 1: We train a classifier for robust deepfakes detection with an auxiliary re-synthesizer trained on real images which are comprised of several tasks for modeling real image distributions, isolating fake images and extracting robust features in unknown scenarios.

Malicious individuals may rely on the above techniques to alter or create the media, and spread misleading information, which will cause unpredictable results. Therefore, several precautions of misuse of the techniques are developed, including targeting the source of forgery [Zhang *et al.*, 2020; Yu *et al.*, 2019b; Yu *et al.*, 2021b; Yu *et al.*, 2021c] and detection of forgery, which provides people warning messages to trust the media suitably.

In particular, detecting fake images has received tremendous attention, because images are the ubiquitous media and appear widely in various platforms, such as social networks, advertisement, etc. Besides, detecting fake images is the backbone of many alarm systems working in more sophisticated cases, such as videos. Recent fake detection techniques mainly rely on local regions artifacts [Yu *et al.*, 2019b; Masi *et al.*, 2020; Chai *et al.*, 2020], global textures [Liu *et al.*, 2020] or rely on a mismatch in the generated frequency distribution [Durall *et al.*, 2019; Zhang *et al.*, 2019; Frank *et al.*, 2020]. A CNN classifier is typically trained on the extracted features to perform binary classification for fake detection. However, recent work has shown that such low-level features are rather easy to recognize [Wang *et al.*, 2020] and can be effectively concealed [Durall *et al.*, 2020; Jung and Keuper, 2021]. Due to the steady improvement of generative models and the constantly narrowing gap between real and fake images, this appears not to yield in a reliable and sustainable approach to distinguish real and fake images. Therefore, it motivates us to seek different, diverse, or at least complementary approaches for robust generated images de-

tection targeting potential unknown configurations.

In order to overcome the issue of excessive reliance on simplistic frequency and low-level artifacts, we propose a novel feature representation for generated images detection. We achieve this by processing *both* real *and* fake images with a generator that in turn induces similar frequency artifacts to *both* images while distinct *residuals*, as sketched in Figure 1. The generator is trained with real images to perform several synthesis tasks. We aim to complete information for real images from some sketched information, such as colorization, denoising and super-resolution. The frequency artifacts become non-discriminative features and will - as we show - not be used for the detection. Instead, the proposed detection mechanism leverages the features of multi-stage reconstruction errors w.r.t. the re-synthesis model. It turns out to be remarkably effective to distinguish real and fake images - which in addition is more generalized across different GAN techniques and more robust against a variety of perturbations that try to conceal fakes.

We highlight the contributions and novelty of our work as follows:(1) We present a novel feature representation for fake image detection based on re-synthesis, which is based on a super-resolution pipeline that learns a detector agnostic to the previously-used simplistic frequency features. (2) We validate the improvements of our method in terms of detection accuracy, generalization, and robustness, compared to prior work in a diverse range of settings.

2 Related Work

2.1 Generative Adversarial Networks (GANs)

GANs [Goodfellow *et al.*, 2014] have achieved tremendous success in modeling data distributions. The breakthroughs mainly come from the improvements of training strategies [Karras *et al.*, 2018] or model architectures [Karras *et al.*, 2019; Karras *et al.*, 2020; Yu *et al.*, 2021a]. The current state-of-the-art GANs are capable of producing high-resolution images with realistic details, which make it rather challenging for human eyes to distinguish generated and real images apart. Specifically, we study the problem of detecting deepfakes with the state of the art GANs: ProGAN [Karras *et al.*, 2018], StarGAN2 [Choi *et al.*, 2020], StyleGAN [Karras *et al.*, 2019], StyleGAN2 [Karras *et al.*, 2020].

2.2 Low-Level Artifacts Produced by GANs

GANs have been significantly improved in recent years and are able to synthesize high fidelity images fooling human eyes. However, there persist some problems in GANs revealing the differences between generated distributions and real ones because of commonly-used up-convolution (or deconvolution) operation [Durall *et al.*, 2020], which maps low-resolution tensors to high-resolution ones. Yet these problems are never long-lasting compared to the steady improvement of GANs. For example, spectral regularization is proposed [Durall *et al.*, 2020] to close the gap in the spectral domain. Recently, [Jung and Keuper, 2021] learn an additional discriminator with spectrum inputs with adversarial training, and the frequency gap of fake images is reduced further. Hence, it is not sustainable to establish fake detection mechanisms based on the known problems of GANs - these problems are also known to malicious individuals and can be sidestepped along with the steady development of improved GANs. That motivates us to propose a novel mechanism for detecting fake images, which is not reliant on such low-level artifacts.

2.3 Fake Detection with Spatial Analysis

Because of the emerging risks of fake information explosion, fake image detection has become an increasingly crucial and prevalent topic [Marra *et al.*, 2018; Nataraj *et al.*, 2019; Rössler *et al.*, 2019; Wang *et al.*, 2020; Marra *et al.*, 2019b]. Recent work analyzes different low-level visual pattern representations so as to attribute images into real or fake [Marra *et al.*, 2019a; Yu *et al.*, 2019b; Liu *et al.*, 2020; Chai *et al.*, 2020]. First, [Yu *et al.*, 2019b; Marra *et al.*, 2019a] validate that GAN training naturally leaves a unique fingerprint for each model, which serves as a visual cue for fake detection. Furthermore, [Liu *et al.*, 2020] design a network to induce texture representations using a Gram matrix, and validate that global textures at different levels of a CNN are effective cues for fake detection. Also, Laplacian of Gaussian (LoG) is augmented along with images to foster fake image and video detection [Masi *et al.*, 2020]. Last, according to a patch-level prediction from different stages of a CNN, it has been shown that hair and background are the most informative areas for detecting fake facial images [Chai *et al.*, 2020], which may help detection across various data distribution. In this paper, we compare to GAN fingerprint techniques [Marra *et al.*, 2019a; Yu *et al.*, 2019b] as representatives of this approach, and show improved performance.

2.4 Fake Detection with Frequency Analysis

Frequency analysis has a long history and broad applications in image processing. Several recent methods based on analyzing frequency patterns of images are adapted to fake detection. A simple yet effective method based on azimuthally-averaged spectrum magnitude and SVM is proposed [Durall *et al.*, 2019]. The 2d-FFT magnitudes serve as input features for CNN binary classification [Zhang *et al.*, 2019; Wang *et al.*, 2020]. In a similar spirit, 2d-DCT is also studied as CNN input features [Frank *et al.*, 2020] and demonstrates improved detection results compared to image-based method [Yu *et al.*, 2019b]. To the best of our knowledge, the most recent state-of-the-art detector leverages global and local 2D DCT features [Qian *et al.*, 2020] and further validates the effectiveness of frequency analysis in terms of detecting fake images. In this paper, we compare to azimuthally-averaged spectrum, 2d-FFT, and 2d-DCT as representatives of this approach, and show improved performance.

3 Detection by Re-Synthesis

The key goal of our work is to propose a more robust feature presentation for fake image detection, which should be generalized enough across different fake sources, robust enough against image perturbations, and more importantly, not reliant on low-level artifacts naturally induced by generative models.

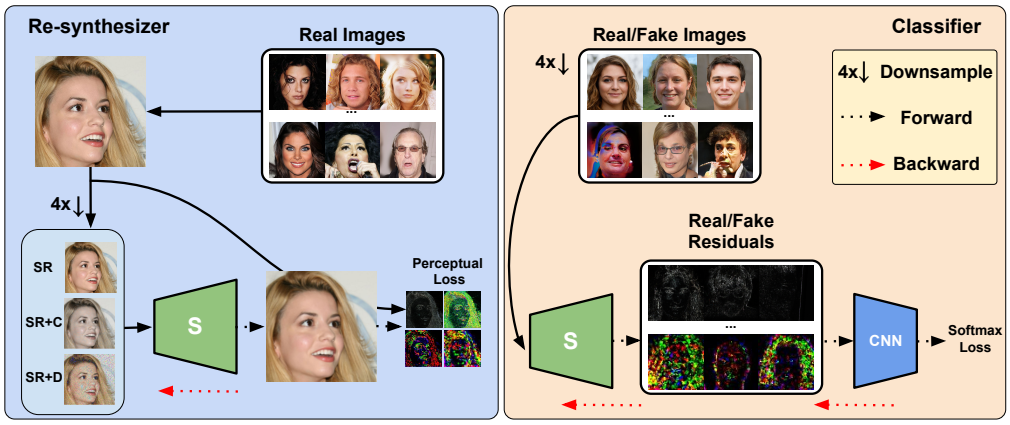

Figure 2: The diagram of our detection pipeline. Our end-to-end model has two components. A classifier is trained with real/fake images. We learn a re-synthesizer with real images only to help extracting robust features and isolating fake images. The synthesizer takes different forms of inputs to capture various visual representations from those tasks, including super-resolution (SR), colorization (C) and denoising (D).

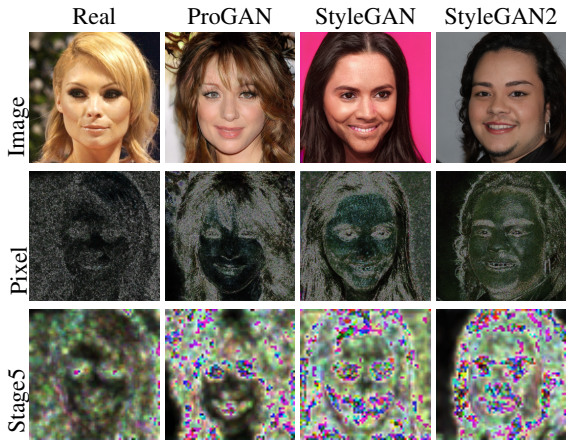

Figure 3: Visualization of hierarchical structural artifacts for different image sources.

To achieve this, we propose a novel method that is explicitly designed to re-synthesize testing images according to several auxiliary tasks to model the distribution for real images, for the purpose of robust fake image detection. Our model has two components, a *re-synthesizer* $\mathbf{S}$ and a *classifier* $\mathbf{C}$, and they are jointly trained end to end. The re-synthesizer aims to distinct real/fake images that it is only learned with real images to provide different residuals from fake ones. Besides, the re-synthesizer allows us to incorporate a variety of noisy patterns to avoid overfitting on the representations from it for the robust detection in unknown scenarios. As an instance, our method incorporates a super-resolution model, which aims to predict the high-frequency information from low-resolution inputs. Owing to the differences w.r.t high-frequency information between real and fake images, this allows us to build a fake image detector based on the residual errors (i.e., reconstruction errors) formalized by this super-resolution model. Further, we take different forms of inputs for the super-resolution model to capture richer and more robust visual representations from real images for improving the robustness of detection.

3.1 Re-Synthesis Residuals as Structural Artifacts

We train a re-synthesizer $\mathbf{S}$ on real images, which takes a downsampled version to reconstruct the original image, and regard the structural reconstruction error maps as features for classification. Particularly, $\mathbf{S}$ is trained on real images only, and then we are able to show different residual distributions for real and fake images.

Mathematically, given a dataset $\mathcal{D}^+$ representing real images, we first train a super-resolution model $\mathbf{\Phi}$ on $\mathcal{D}^+$, which is formulated as a regression task with the loss function

$$L = \|X - \mathbf{\Phi}(\mathbf{\Omega}(X_{\downarrow}))\|_1, \tag{1}$$

where $\mathbf{\Omega}$ could be an image degeneration operation, $X \in \mathcal{D}^+$ and $X_{\downarrow}$ is a downsampled version of X. As a result, our re-synthesizer $\mathbf{S}(\cdot) = \mathbf{\Phi}(\mathbf{\Omega}(\cdot))$. In this work, $\mathbf{\Omega} = \{I, G, N\}$, referring to the indentity, grayscaling and noising operations respectively.

After training the super-resolution model, we apply the structural artifact with downsampled images $|X - \mathbf{\Phi}(X_{\downarrow})|$ as the feature for fake detection. We collect another dataset of fake images from a generator, denoted as $\mathcal{D}^-$. A fake image detector is then trained on $\mathcal{D}^+ \cup \mathcal{D}^-$. We formulate the detector as a neural network classifier $\mathbf{C}(\cdot)$ trained with softmax loss. In testing, given a query image X^*, the detection decision is formulated as

$$\mathbf{C}(|X^* - \mathbf{\Phi}(X^*_{\downarrow})|). \tag{2}$$

3.2 Hierarchical Artifacts via Perceptual Loss

Despite the success of low-level artifacts in fake detection [Yu *et al.*, 2019b; Masi *et al.*, 2020], high-level information remains unexplored and should be equally effective, such as semantic parts, global coherence, etc. Perceptual loss [Johnson *et al.*, 2016], on the other hand, evaluates the difference between two examples w.r.t CNN activations at various layers, which correspond to hierarchical representations of visual information. Consequently, in order to boost our detection pipeline with high-level information, we take advantage of perceptual loss from a pretrained network to train our

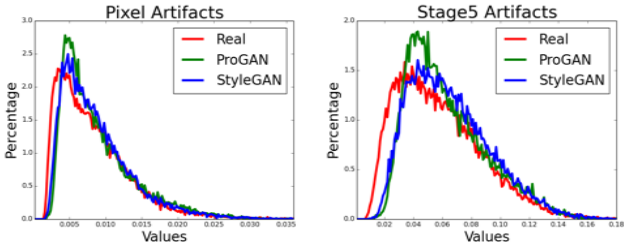

Figure 4: Histograms of the spatially-average amounts of hierarchical structural artifacts on CelebA-HQ. We observe clear margins between distributions of real and fake (ProGAN or StyleGAN).

super-resolution model, instead of only considering the reconstruction errors in pixels.

Let $\mathbf{\Theta}$ be a pretrained network, and $\mathbf{\Theta}_i(\cdot)$ be the operation to extract features in the i-th stage of a total of n stages. The loss function with perceptions and image degeneration is formulated as

$$L = \alpha_0 \|X - \mathbf{\Phi}(\mathbf{\Omega}(X_\downarrow))\|_1 + \sum_{i=1}^{n} \alpha_i \|\mathbf{\Theta}_i(X) - \mathbf{\Theta}_i(\mathbf{\Phi}(\mathbf{\Omega}(X_\downarrow)))\|_1, \quad (3)$$

where $\alpha_0, \alpha_1, ..., \alpha_n$ are loss weights to control the importance of corresponding feature stages during training. After training the super-resolution model, we leverage the ℓ_1 residual map at each stage, as the feature for fake detection. Notably, we can detect fake images by a single classifier or by a combination of classifiers at different stages. Let $\{\mathbf{C}_0(\cdot), \mathbf{C}_1(\cdot), ..., \mathbf{C}_n(\cdot)\}$ be the set of classifiers, where we define $\mathbf{C}_0(\cdot)$ is the pixel-level classifier, and others are the classifiers trained on artifacts at different stages from the perceptual loss. The final decision for input image X^* is computed with weights $\beta_0, \beta_1, ..., \beta_n$ and formulated as

$$\beta_0 \mathbf{C}_0(|X^* - \mathbf{\Phi}(X^*_\downarrow)|) + \sum_{i=1}^{n} \beta_i \mathbf{C}_i(|\mathbf{\Theta}_i(X^*) - \mathbf{\Theta}_i(\mathbf{\Phi}(X^*_\downarrow))|). \quad (4)$$

In Figure 3, we show examples from CelebA-HQ [Karras *et al.*, 2018] and their structural artifacts in the pixel level and in stage5 of the perceptual network. We observe: (1) The magnitudes of artifacts of fake images are larger than those of real images, which lay the foundation to distinguish fake from real. In particular, there are more severe artifacts on hairs, eyes, or mouths, which is consistent with the recent study of generalization of fake detection [Chai *et al.*, 2020]. (2) The artifact structures are distinct between real and fake, where they look more randomly distributed in real images while with stronger patterns in fake images. (3) The stage5 perceptual artifacts are more discriminative than pixel artifacts to attribute fake, even between ProGAN and StyleGAN trained on the same dataset. This results in the potential for cross-GAN fake detection. For quantitative demonstration, in Figure 4 we plot the histograms of spatially-averaged amounts of artifacts in the pixel level and in stage5, and show the clear margins between distributions of real and fake.

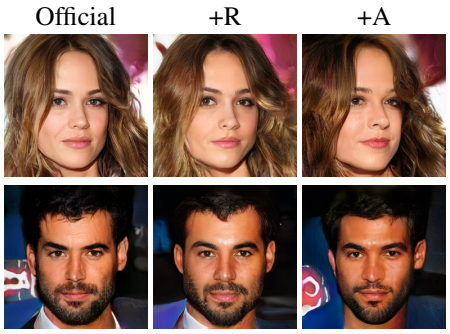

Figure 5: Testing fake image examples from the official ProGAN, as well as the counterparts with regularization (+R) [Durall *et al.*, 2020] and adversarial training (+A) [Jung and Keuper, 2021].

4 Experiments

We present the experimental setup and configuration details from section 4.1 to 4.5. Accordingly, we provide the results and discussion in the rest of this section.

4.1 Datasets

CelebA-HQ is a 1024×1024 facial image dataset released by [Karras *et al.*, 2018]. We choose to necessarily conduct experiments on faces because of the prevalent applications of identity recognition and biometric feature recognition. From this dataset, we prepare 25k/25k real and fake images as the training set, and 2.5k/2.5k images as the testing set.

FFHQ [Karras *et al.*, 2019] provides another 70k 1024×1024 facial images. Compared to CelebA-HQ, it covers vastly more variation on age, ethnicity, eyeglasses, hats, background, etc. To perform cross-domain detection, we test on this dataset all the detectors trained on CelebA-HQ.

LSUN is large-scale scene dataset [Yu *et al.*, 2015]. We select bedroom and church_outdoor classes to perform fake detection beyond human faces. We prepare 50k/50k real and fake images for training, and 10k/10k images for testing.

4.2 Deep Fake Detection Methods

We compare our method to the recently-proposed methods, including PRNU [Marra *et al.*, 2019a], plain image [Yu *et al.*, 2019b], 1d spectrum [Durall *et al.*, 2019], FFT-2d magnitude [Wang *et al.*, 2020; Zhang *et al.*, 2019], DCT-2d [Frank *et al.*, 2020] and GramNet [Liu *et al.*, 2020]. For our models, we report the performance based on the pixel- and stage5-level artifacts using ResNet-50 [He *et al.*, 2016], which capture low- and high-level spatial errors respectively. In addition, we also report the performance of averaging between the two artifacts. In particular, we compare three re-synthesizers in our study: (1) A super-resolution model is used which predicts high-resolution images from downsampled images, namely **SR** for short; (2) A super-resolution model which takes partially gray-scaled images to predict high-resolution color images , referred as **SR+C**. 50% images are performed grayscaling operation. For those images, $10\% \sim 25\%$ pixels are randomly set to gray-scale versions. (3) A super-resolution model which takes noisy images to predict high-

Method	S	ProGAN					StyleGAN					ProGAN→StyleGAN					StyleGAN→ProGAN					Avg
		Raw	+R	+E	+A	+P	Raw	+R	+E	+A	+P	Raw	+R	+E	+A	+P	Raw	+R	+E	+A	+P	
PRNU	/	78.3	57.1	63.5	53.2	51.3	76.5	68.8	75.2	63.0	61.9	47.4	44.8	45.3	44.2	48.9	48.0	55.1	53.6	51.1	53.6	57.3
Image		99.9	58.0	99.9	56.7	**78.8**	99.9	83.9	99.9	72.3	**81.3**	51.6	50.9	51.7.	51.6	51.1	52.8	50.5	54.4	51.5	54.2	67.6
1D Spectrum		97.5	69.6	75.7	70.0	54.9	93.0	49.1	68.2	52.2	54.4	93.0	48.6	71.6	49.6	53.5	97.5	65.0	56.9	66.1	52.3	67.9
FFT-2d		99.9	95.9	81.8	**99.9**	59.8	**100**	90.8	72.0	99.4	57.7	98.9	99.8	63.2	61.1	56.8	77.5	54.6	56.5	76.5	55.5	78.1
DCT-2d		99.9	99.9	99.9	**99.9**	54.4	99.9	99.8	99.9	99.8	56.0	98.6	**99.9**	98.4	81.8	54.2	99.0	95.6	98.6	97.1	55.0	89.2
GramNet		100	77.1	100	77.7	69.0	100	96.3	100	96.3	73.3	64.0	57.3	63.7	50.9	57.1	63.1	56.4	63.8	66.8	56.2	74.6
Ours (Pix)		**100**	99.7	99.9	99.3.	57.5	**100**	70.0	**100**	97.9	57.4	99.8	78.6	**99.9**	98.4	55.8	98.2	99.6	99.2	98.1	55.0	87.0
Ours (Stage5)	SR	99.9	99.9	99.9	**99.9**	71.7	**100**	99.8	**100**	99.8	68.4	99.4	99.5	99.7	99.4	70.7	96.0	97.9	97.1	98.7	67.4	92.5
Ours (Avg)		**100**	**100**	**100**	99.7	64.5	**100**	98.7	**100**	**99.9**	66.7	**100**	97.8	**99.9**	99.8	67.0	**99.5**	**99.9**	99.8	**100**	66.1	92.2
Ours (Pix)		99.9	99.8	99.9	99.4	55.3	**100**	75.1	**100**	98.5	59.5	99.8	80.8	**99.9**	98.4	55.9	99.0	99.5	99.2	98.3	55.1	88.7
Ours (Stage5)	SR+C	99.9	99.9	99.9	**99.9**	72.5	**100**	99.8	**100**	99.7	69.2	99.4	**99.9**	99.7	**99.9**	**72.5**	97.2	98.4	98.6	98.6	**68.0**	**93.7**
Ours (Avg)		**100**	**100**	**100**	99.7	64.5	**100**	99.5	**100**	**99.9**	68.5	**100**	97.8	**99.9**	99.8	68.5	99.3	**99.9**	**100**	**100**	66.1	93.2
Ours (Pix)		**100**	99.8	99.9	99.3.	58.2	**100**	77.3	**100**	99.2	59.9	99.7	79.0	**99.9**	98.2	57.2	99.2	99.8	99.2	98.1	55.0	88.9
Ours (Stage5)	SR+D	99.9	99.9	99.9	**99.9**	73.0	**100**	**99.9**	**100**	99.7	73.4	99.2	99.4	99.8	99.4	69.5	95.2	97.6	97.1	98.2	66.8	93.4
Ours (Avg)		**100**	**100**	**100**	99.7	66.5	**100**	98.7	**100**	**99.9**	68.8	**100**	97.6	**99.9**	99.8	67.2	98.0	**99.9**	99.8	**100**	65.2	93.1

Table 1: Classification accuracy (%) on CelebA-HQ. The models are trained with fake images from either ProGAN or StyleGAN, and tested on a variety of settings including spectrum regularization (+R), spectrum equalization (+E), spectral-aware adversarial training (+A) and an ordered combination of image perturbations (+P). For our models, we apply 3 re-synthesizers including super-resolution (SR), super-resolution combined with colorization (SR+C) and super-resolution combined with denoising (SR+D).

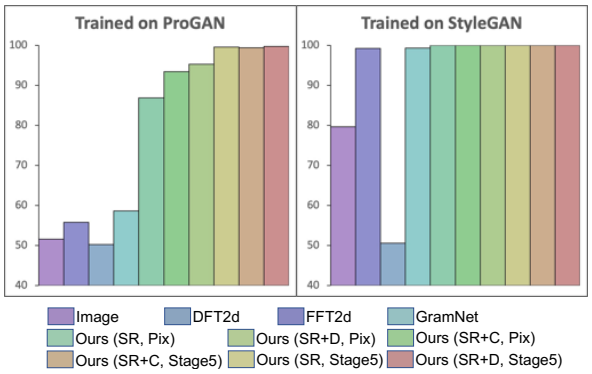

Figure 6: Classification accuracy (%) on StarGAN2 (CelebA-HQ). We employ detectors trained on ProGAN or StyleGAN. The robustness of our models is clearly observed w.r.t cross-domain detection.

resolution clean images, referred as **SR+D**. 50% images are performed noising operation and Gaussian noises with standard deviation 4/255 are applied.

4.3 Robustness against Image Perturbations

Image perturbations are able to alter image details or distributions while preserving the contents of them (e.g., denoising, JPEG compression, etc), which may be used to process fake images and challenge the detectors. In order to compare the robustness of different methods, we test on perturbed fake images that are not seen in training phases. First, we calibrate the frequency distributions of generated images using the equalization operation [Durall *et al.*, 2020]. We regard this operation as post-processing to reduce the frequency distribution gap between real and fake images. Second, we follow the protocol in the previous work [Yu *et al.*, 2019b] to perturb testing images with an ordered combination of JPEG compression, blurring, cropping, and adding Gaussian noise.

4.4 Generators

We generate fake images from ProGAN [Karras *et al.*, 2018], StarGAN2 [Choi *et al.*, 2020], StyleGAN [Karras *et al.*, 2019], and StyleGAN2 [Karras *et al.*, 2020]. To overcome the known frequency artifacts of up-convolution operations in the generators and provide unknown detection scenarios, we apply the spectral regularization (+R) [Durall *et al.*, 2020] and spectral-aware adversarial training (+A) [Jung and Keuper, 2021] to finetune released GANs and Figure 5 shows some examples. Besides, we also apply spectrum equalization (+E), which is similar to the regularization, but we process the fake images as a postprocessing.

4.5 Implementation Details

We train the super-resolution model on the real images from CelebA-HQ or LSUN. We build a $4\times$ super-resolution model using [Zhang *et al.*, 2018] supervised by the ell_1 pixel loss plus VGG-based perceptual loss [Johnson *et al.*, 2016; Wang *et al.*, 2018]. We set the loss weight for each feature from pixel to stage 5 as [1, 1/32, 1/16, 1/8, 1/4, 1]. Second, we train the detectors on pixel artifacts and stage5 artifacts from the VGG network. For each detector, we train a ResNet-50 [He *et al.*, 2016] from scratch for 20 epochs using the SGD optimizer with momentum. The initial learning rate is 0.01 and we reduce it to 0.001 at the 10th epoch.

4.6 Results on CelebA-HQ

We conduct experiments on CelebA-HQ and compare our method to the baselines, as listed in Table 1. In specific, our model (Avg) leverages the hierarchical structural artifacts in a combination mode as described in Eq. (4), where we set (1/2 , 1/2) as the weights for combining the final scores of the pixel and stage 5 artifacts. In Table 1, we present the results including the training and testing images are from the same generator and the images are from the different generator. For each part, we not only test on the raw images, but also test on the different challenges of spectrum regularization (+R), equalization (+E), adversarial training (+A) and a combination of perturbations (+P), as discussed before.

Method	S	ProGAN			StyleGAN			Avg
		SG	SG2[1]	SG2[2]	SG	SG2[1]	SG2[2]	
PRNU	/	46.3	44.5	46.1	63.2	53.7	43.5	45.4
Image	/	49.8	45.2	44.9	50.0	48.0	49.2	47.9
1D Spectrum	/	51.7	54.5	51.4	50.9	61.3	53.7	53.9
FFT-2d	/	50.7	58.4	72.0	56.1	94.5	95.2	71.1
DCT-2d	/	**87.3**	62.4	67.3	**88.8**	93.8	93.6	82.2
GramNet	/	50.1	45.6	45.7	66.2	46.8	48.5	50.5
Ours (Pix)	SR	70.6	81.2	91.3	60.2	95.6	96.2	82.5
Ours (Stage5)	SR	83.0	54.0	55.1	86.2	71.3	68.3	69.7
Ours (Avg)	SR	79.1	62.1	70.9	77.8	89.7	90.0	78.3
Ours (Pix)	SR+C	71.6	**93.4**	93.8	64.6	97.2	96.9	86.3
Ours (Stage5)	SR+C	83.1	54.1	55.8	85.5	71.5	68.4	69.7
Ours (Avg)	SR+C	81.2	88.2	88.5	80.3	91.5	92.0	87.0
Ours (Pix)	SR+D	71.0	**93.4**	**94.1**	65.4	**97.4**	**97.0**	86.4
Ours (Stage5)	SR+D	81.8	53.5	54.6	83.7	66.8	65.7	67.7
Ours (Avg)	SR+D	78.8	88.8	89.6	79.7	93.9	93.4	**87.4**

Table 2: Classification accuracy (%) w.r.t cross-domain detection on FFHQ. Classification models are trained on CelebA-HQ with ProGAN or StyleGAN. Fake images from StyleGAN and StyleGAN2 trained on FFHQ are used for testing, abbreviated as SG and SG2. For StyleGAN2, we generate images with psi value of 0.5 or 1.0, referred as $SG2^1$ and $SG2^2$.

First, we observe our detectors achieve remarkable results in average. We emphasize: (1) Spectrum-based methods [Durall *et al.*, 2019; Zhang *et al.*, 2019] achieves decent performance and robustness in the cross-GAN settings. For example, the performance of [Durall *et al.*, 2019] and [Zhang *et al.*, 2019] deteriorates only by 10% on average in the cross-GAN settings. (2) DCT-2d-based method is robust across domains and robust against processing on frequency. However, it is sensitive to the image perturbations, deteriorating the performance severely. (3) In contrast, our model with any of the features achieves better results compared to the baselines. Especially the deterioration is imperceptible when applying spectral processing or testing cross domains. (4) Introducing colorization or denoising into the re-synthesis helps learning more robust features and achieving more favorable results.

In addition to our advantageous performance, we also point out a few insightful discoveries as follows: (1) Our stage5-based detector obtains the best accuracy in the cross-domain settings with image perturbations, which validates the robustness of leveraging high-level information. (2) Our average detector achieves better performance than our other two detectors in many tests, indicating the beneficial synergy between our pixel artifacts and stage 5 artifacts. We reason this as the synergy effect between pixel and stage artifacts. Therefore, we suggest considering more about high-level cues for fake detection, instead of looking at the local only.

Additionally, we also test the above detectors on recent proposed StarGAN2, where the results are visualized in Figure 6. We highlight several points: (1) Most competing detectors fail to recognize StarGAN2 generated images owing to the domain shift. Even though DCT2d detector perform well in Table 1, it hardly recognizes the fakeness from StarGAN2 images. (2) All of our detectors with different synthesis modules and visual features achieve accuracy higher than 85% and most of them are close to 100%. Besides, the improved results are achieved for ProGAN after employing colorization or denoising in the re-synthesis, which shows the effectiveness of introducing noises into the re-synthesizer.

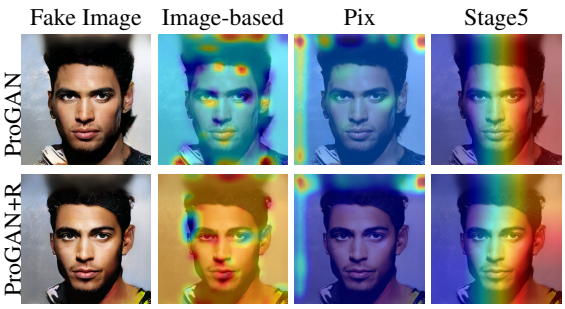

Figure 7: Visualization of class activation maps for recognizing as fake images in CelebA-HQ.

To further demonstrate the robustness of our detectors, we visualize the class activation maps (CAMs) for our methods (SR) and previous detectors [Yu *et al.*, 2019b] in Figure 7. From this figure, we are able to observe the CAMs of our approach are quite stable for similar input fake images, even though spectral regularization [Durall *et al.*, 2020] is applied to reduce the spectrum distrbution between real and fake images. In contrast, the image based detector [Yu *et al.*, 2019b] is sensitive to the changes and outputs an opposite highlighted region. Therefore, we reason performance reduction for previous approaches to their sensitivities.

4.7 Results on FFHQ

We conduct experiments for cross-domain detection with FFHQ, where the results are listed in Table 2. In the experiments, we do not train additional detectors; instead directly test the models on novel data, which are trained on CelebA-HQ. We apply the real images from FFHQ and fake images from StyleGAN and StyleGAN2 trained on FFHQ. Because this setting already challenges most detectors, we do not employ additional perturbations as CelebA-HQ. From Table 2 we observe: (1) Our model using pixel inputs achieves the best performance, and outperforms the DCT-2d based approach when StyleGAN2 is tested. (2) Our stage5-based models achieve comparable performance to DCT-2d and are better than our pixel-based when StyleGAN is tested, which further shows the complementary capability of low- and high-level artifacts to deal with wider detection scenarios. (3) We can observe clear improvements when colorization or denoising is applied, indicating more robust features are obtained and thus achieving better results in cross-datasets detection. We owe this observation to the avoid of overfitting on the training data by the noisy inputs of our synthesizers.

4.8 Results on LSUN

We use official StyleGAN to generate images of bedroom and StyleGAN2 to generate images of church_outdoor. In Table 3, we test all the models on raw images and those combined with perturbations. We observe our method reaches to competing results compared with other state of the arts. In particular, our detector using stage 5 artifacts is the most stable against perturbations, which shows the necessity of involving high-level information. For example, our SR+C and

Method	S	Bedroom		Church_Outdoor		Avg
		Raw	+P	Raw	+P	
Image	/	**100**	54.2	**99.9**	57.7	78.0
FFT-2d		99.6	61.5	99.8	57.5	79.6
DCT-2d		99.9	59.5	**99.9**	60.5	80.0
GramNet		99.9	73.7	99.9	68.7	85.5
Ours (Pix)	SR	99.9	56.6	99.8	58.4	78.7
Ours (Stage5)		99.6	**76.5**	99.5	75.2	87.7
Ours (Avg)		99.9	65.1	**99.9**	69.9	83.7
Ours (Pix)	SR+C	99.9	57.5	99.8	56.4	78.4
Ours (Stage5)		99.6	76.1	99.0	**96.0**	**92.7**
Ours (Avg)		99.9	71.0	99.8	73.0	86.0
Ours (Pix)	SR+D	99.9	56.9	99.7	57.0	78.4
Ours (Stage5)		99.2	75.9	98.8	95.9	92.2
Ours (Avg)		99.9	71.4	99.8	72.9	86.0

Table 3: Classification accuracy (%) on LSUN datasets. We test on the fake images from official released models (Raw) and employ a combination of perturbations on them (+P).

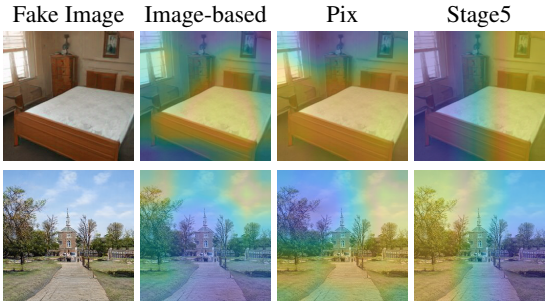

Figure 8: Visualization of class activation maps for recognizing as fake images in LSUN.

SR+D detectors based on stage 5 achieve more than 95% accuracy on church_outdoor with perturbations while others are less than 70%. We conclude the proposed method is able to cope with diverse challenges using multi-level features and our re-synthesis module. It sheds the light on the new way for fake image detection, more advantageous to only analyzing the local patterns [Marra *et al.*, 2019a; Yu *et al.*, 2019b] or frequency distributions [Frank *et al.*, 2020; Zhang *et al.*, 2019]. Additionally, we show the class activation maps for LSUN in Figure 8. Unlike facial images, our CAMs are more flexible in highlighted shapes and locations for highly diverse scene images in LSUN.

5 Conclusion

Due to the unsustainable reliance of frequency based deepfake detectors, we show how these methods deteriorate when such artifacts are suppressed. Based on this insight, we present a novel feature representation for detecting Deepfakes. Instead of the limited focus on low-level local artifact characteristics, we reason that different levels of information can help detect fake images with beneficial synergy. The hierarchical artifacts from a re-synthesizer are evidenced to boost the performance, generalization, and robustness of a downstream detector. These have been validated with high-resolution Deepfakes created from state-of-the-art GANs. Further, deepfakes detection is still an open problem because many issues are unresolved. Attacks can reverse the procedure of classification to fool a detector, therefore, we believe our solution provides a promising solution that our features are from a parameterized model which can be watermarked, and thus increases the challenges of attacks on our classifier. In the end, we conclude deepfakes detection requires more than analysis in addition to low-level details, but also on higher-level visual cues which has the potential to lead to more sustainable detection schemes.

Acknowledgements

Ning Yu is partially supported by the Twitch Research Fellowship.

References

[Chai *et al.*, 2020] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020.

[Choi *et al.*, 2018] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[Choi *et al.*, 2020] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020.

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

[Durall *et al.*, 2020] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 2020.

[Frank *et al.*, 2020] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[Jung and Keuper, 2021] Steffen Jung and Margret Keuper. Spectral distribution aware image generation. *arXiv preprint arXiv:2012.03110*, 2021.

[Karras *et al.*, 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[Karras *et al.*, 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

[Liu *et al.*, 2020] Zhengzhe Liu, Xiaojuan Qi, Jiaya Jia, and Philip Torr. Global texture enhancement for fake face detection in the wild. In *CVPR*, 2020.

[Marra *et al.*, 2018] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *MIPR*, 2018.

[Marra *et al.*, 2019a] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *MIPR*, 2019.

[Marra *et al.*, 2019b] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. A full-image full-resolution end-to-end-trainable cnn framework for image forgery detection. *arXiv preprint arXiv:1909.06751*, 2019.

[Masi *et al.*, 2020] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, 2020.

[Nataraj *et al.*, 2019] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020.

[Radford *et al.*, 2016] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

[Rössler *et al.*, 2019] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

[Thies *et al.*, 2016] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[Wang *et al.*, 2018] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

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

[Wang *et al.*, 2021] Hui-Po Wang, Ning Yu, and Mario Fritz. Hijack-gan: Unintended-use of pretrained, black-box gans. *arXiv preprint arXiv:2011.14107*, 2021.

[Yu *et al.*, 2015] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[Yu *et al.*, 2019a] Ning Yu, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, and Michal Lukac. Texture mixer: A network for controllable synthesis and interpolation of texture. In *CVPR*, 2019.

[Yu *et al.*, 2019b] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *ICCV*, 2019.

[Yu *et al.*, 2020] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In *ECCV*, 2020.

[Yu *et al.*, 2021a] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry Davis, and Mario Fritz. Dual contrastive loss and attention for gans. *arXiv preprint arXiv:2103.16748*, 2021.

[Yu *et al.*, 2021b] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. *arXiv preprint arXiv:2007.08457*, 2021.

[Yu *et al.*, 2021c] Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. *arXiv preprint arXiv:2012.08726*, 2021.

[Zhang *et al.*, 2018] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018.

[Zhang *et al.*, 2019] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *WIFS*, 2019.

[Zhang *et al.*, 2020] Baiwu Zhang, Jin Peng Zhou, Ilia Shumailov, and Nicolas Papernot. Not my deepfake: Towards plausible deniability for machine-generated media. *arXiv preprint arXiv:2008.09194*, 2020.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.