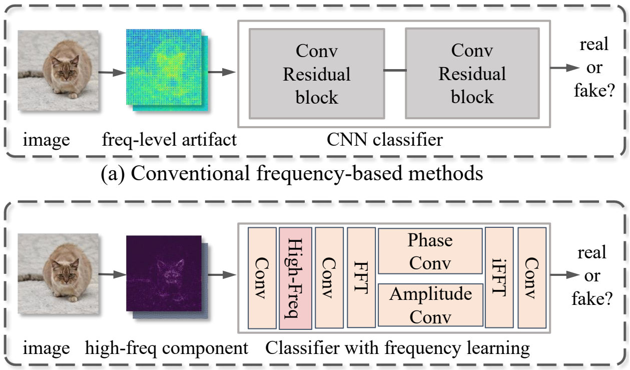
Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning

Chuangchuang Tan[1,2], Yao Zhao[1,2*], Shikui Wei[1,2], Guanghua Gu[3,4], Ping Liu[5], Yunchao Wei[1,2]

[1]Institute of Information Science, Beijing Jiaotong University
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology
[3]School of Information Science and Engineering, Yanshan University
[4]Hebei Key Laboratory of Information Transmission and Signal Processing
[5]Center for Frontier AI Research, IHPC, A*STAR, Singapore
{tanchuangchuang, yzhao, shkwei}@bjtu.edu.cn, guguanghua@ysu.edu.cn, pino.pingliu@gmail.com,
wychao1987@gmail.com

Abstract

This research addresses the challenge of developing a universal deepfake detector that can effectively identify unseen deepfake images despite limited training data. Existing frequency-based paradigms have relied on frequency-level artifacts introduced during the up-sampling in GAN pipelines to detect forgeries. However, the rapid advancements in synthesis technology have led to specific artifacts for each generation model. Consequently, these detectors have exhibited a lack of proficiency in learning the frequency domain and tend to overfit to the artifacts present in the training data, leading to suboptimal performance on unseen sources. To address this issue, we introduce a novel frequency-aware approach called FreqNet, centered around frequency domain learning, specifically designed to enhance the generalizability of deepfake detectors. Our method forces the detector to continuously focus on high-frequency information, exploiting high-frequency representation of features across spatial and channel dimensions. Additionally, we incorporate a straightforward frequency domain learning module to learn source-agnostic features. It involves convolutional layers applied to both the phase spectrum and amplitude spectrum between the Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (iFFT). Extensive experimentation involving 17 GANs demonstrates the effectiveness of our proposed method, showcasing state-of-the-art performance (+9.8%) while requiring fewer parameters. The code is available at https://github.com/chuangchuangtan/FreqNet-DeepfakeDetection.



(a) Conventional frequency-based methods

(b) FreqNet: Frequency space learning Network

Figure 1: Frequency space learning network. (a) The traditional studies are usually limited to developing frequency-level artifacts. (b) Distinguishing itself from prior frequency-based research, our approach shifts its focus to the frequency-related attributes of the features within the detector.

Introduction

The proliferation of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014; Karras et al. 2018, 2019) has significantly simplified the generation of lifelike synthetic images, resulting in an alarming surge in the prevalence of forgeries that are virtually indistinguishable from authentic images to the human visual system. This escalating trend poses potential, unpredictable societal repercussions. In response, a multitude of deepfake detection mechanisms have been conceived (Frank et al. 2020; Li et al. 2021), with specific emphasis on detecting facial forgeries. However, the majority of existing forgery detection techniques suffer from a fundamental limitation: they are constrained to the same domain during both their training and evaluation phases. This limitation severely hampers their ability to generalize effectively to unseen domains, such as those involving unfamiliar generation models or novel categories.

The pursuit of Generalizable Deepfake Detection strives to create a universal detector capable of effectively identifying deepfake images even when faced with limited training data, a necessity given the continual emergence of increasingly sophisticated synthesis technologies. Recently, prior investigations (Frank et al. 2020; Durall et al. 2020) have substantiated the efficacy of frequency artifacts in the realm of deepfake detection, notably in the context of facial detection (Qian et al. 2020; Luo et al. 2021). The findings of (Frank et al. 2020; Durall et al. 2020) have unveiled the presence of significant artifacts within the frequency domain

[*] Corresponding author

stemming from the upsampling operations within GAN architectures. Consequently, this revelation has spurred the development of numerous frequency-based methods aimed at detecting images with pronounced frequency-related characteristics.

Nonetheless, owing to the extraordinary advancements in synthesis technology, an increasing array of distinctive frequency-level artifact representations have emerged. In Figure 2, we present the mean Fast Fourier Transform (FFT) (Cooley et al. 1969) spectrum of images sampled from various sources. This mean spectrum computation involves averaging over 2,000 images, following the methodology detailed in (Frank et al. 2020). Notably, the results exhibit discernible differences in artifact characteristics across different GANs, further accentuated by variations within the same GAN architecture when training on dissimilar datasets. The frequency attributes of images indeed possess the capacity to unveil distinctions between real and generated images. However, they exhibit limitations in terms of generalization across a diverse range of sources.

To surmount this challenge, the primary approach entails the development of a robust classifier specifically designed for frequency representations. (Jeong et al. 2022c), in addressing this issue, chooses to disregard frequency-level artifacts in images by devising a frequency-level perturbation generator. However, this solution introduces complexity and incurs considerable computational expenses. In the realm of deepfake detection, it is imperative to factor in the notion of the detector learning within the frequency domain. We deliberately refrain from directly utilizing the frequency information as the artifact representation to train a CNN classifier. Instead, we strategically compel the detector to acquire its understanding within the frequency space. This nuanced strategy holds the key to achieving a more generalizable deepfake detection framework.

Drawing upon intuitive insights, we introduce a novel and lightweight approach named FreqNet, which integrates frequency domain learning into a lightweight CNN classifier, aimed at enhancing the generalization capabilities of the detector. Diverging from existing frequency-based studies that predominantly focus on the frequency domain of images, the principal innovation of our FreqNet method resides in the simultaneous development of both frequency-domain information derived from images and the features extracted by CNN model. Specifically, our approach introduces two critical modules: the high-frequency representation and the frequency convolutional layer, each meticulously designed to facilitate frequency space learning. The first module serves to compel the detector to consistently prioritize high-frequency information, augmenting its sensitivity to significant details. Additionally, to capture broader forgery indicators within the frequency domain, we incorporate a frequency convolutional layer, which effectively diminishes the reliance on source-specific characteristics. By virtue of this frequency domain-based learning strategy, our proposed FreqNet remarkably extends its generalizability to previously unseen sources with few parameters.

To comprehensively assess the extent of its generalization capabilities, we conduct extensive simulations using an extensive image database generated by 17 distinct models [1]. Despite its modest scale, FreqNet, boasting 1.9 million parameters, significantly outperforms the current state-of-the-art model boasting 304 million parameters, demonstrating a remarkable improvement of 9.8%.

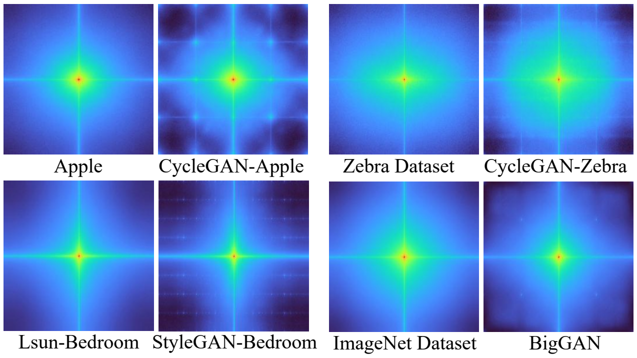


Figure 2: Frequency analysis on various sources. This mean FFT spectrum computation involves averaging over 2,000 images, following the methodology detailed in (Frank et al. 2020).

Our paper makes the following contributions:

- We present a novel frequency space learning network, FreqNet, to achieve generalizable deepfake detection. Our approach strategically incorporates frequency domain learning within a CNN classifier, resulting in a significant enhancement of the detector's ability to generalize across diverse scenarios.
- We utilize convolutional layers on both the phase spectrum and the amplitude spectrum as a deliberate strategy to capture broader forgery indicators within the frequency domain. This allows us to enhance the detector's capability to identify a broader range of artifacts.
- The proposed lightweight FreqNet, consisting of a mere 1.9 million parameters, impressively outperforms the current state-of-the-art model featuring 304 million parameters, attributed mainly to its utilization of frequency domain learning.

Related Work

In this section, we present a concise survey of deepfake detection methodologies, categorizing them into two primary classes: image-based detection and frequency-based detection.

Image-based Deepfake Detection

There has been a significant effort in the field of forgery detection, with many studies focusing on leveraging spatial information from images. Rossler *et al.*(Rossler et al. 2019) utilize images to train the Xception (Chollet 2017) to achieve fake face image detection. Other image-based detection methods have developed specific artifact detection in

[1] ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, Deepfake, AttGAN, BEGAN, CramerGAN, InfoMaxGAN, MMDGAN, RelGAN, S3GAN, SNGAN, STGAN

distinct facial regions, such as eyes(Li et al. 2018) and lips (Haliassos et al. 2021). Chai *et al.*(Chai et al. 2020) adopt limited receptive fields to identify patches that render images detectable, highlighting the importance of specific local features. With the emergence of deepfake technology, efforts are being made to enhance the generalization ability of detectors, particularly for unseen data. Yu *et al.*(Yu et al. 2020) introduce artifacts from the camera imaging process. Furthermore, various methods (Wang et al. 2020, 2021; Chen et al. 2022; Cao et al. 2022; He et al. 2021; Shiohara et al. 2022) aim to enrich the diversity of training data, employing techniques such as data augmentation, adversarial training, reconstruction, and blending images. CDDB (Li et al. 2023) adopts incremental learning to achieve continual deepfake detection. Notably, recent works by Ojha *et al.*(Ojha et al. 2023) and Tan *et al.*(Tan et al. 2023) employ the feature map and gradients as the general representation, respectively.

Frequency-based Deepfake Detection

The research by (Frank et al. 2020; Durall et al. 2020), which highlights the effectiveness of frequency artifacts in the domain of deepfake detection, has significantly influenced subsequent studies. These findings have led many image forgery detectors to shift their attention toward capturing unique patterns within the frequency domain. The work by Masi et al. (Masi et al. 2020) stands out as it meticulously investigates artifacts present in both the color space and the frequency domains, while F^3-Net (Qian et al. 2020) suggests using the discrepancy of frequency statistics between real and forged images as a means to differentiate face image manipulations. An adaptive frequency features learning is designed by FDFL (Li et al. 2021) to mine subtle artifacts from the frequency domain, enabling the detection of forged images. Moreover, ADD (Woo et al. 2022) incorporates two meticulously designed distillation modules to emphasize the significance of frequency information, incorporating frequency attention distillation and multi-view attention distillation. Recently, frequency-based studies have been proposed for generalized detection. BiHPF (Jeong et al. 2022a) emphasizes amplifying artifact magnitudes through the utilization of dual high-pass filters. The FreGAN model, introduced by (Jeong et al. 2022c) ingeniously mitigates the impact of frequency-level artifacts through the deployment of frequency-level perturbation maps. (Wang et al. 2023) introduces dynamic graph learning to exploit the relation-aware features in spatial and frequency domains.

Methodology

In this section, we present our FreqNet technique, a universal deepfake detection method for generalizable deepfake detection. We illustrate the overall architecture of FreqNet in Figure 3, leveraging frequency domain learning to mitigate source-specific dependencies.

Problem Definition

Our primary focus lies within the realm of Generalizable Deepfake Detection. Our objective is to construct a universal detector capable of accurately identifying deepfake images even when confronted with constrained training sources. In this context, let us consider a real-world image scenario denoted as X sampled from n different sources:

$$\begin{aligned} &X = \{X_1, X_2, ..., X_i, ..., X_n\}, \\ &X_i = \{x_j^i, y_j\}_{j=1}^{N_i}, \end{aligned} \tag{1}$$

where N_i represents the number of images originating from the ith source X_i, x_j^i is the jth image of X_i. Each image is labeled with y, indicating whether it belongs to the category of ”real”($y = 0$) or ”fake” ($y = 1$). Here we train a binary classifier $D(\cdot)$, utilizing the training source X_i:

$$D^i = \arg\min_{\theta} \; l(D(X_i; \theta), \; y), \tag{2}$$

where $l()$ denote the loss function. Our overarching goal is to design a detector that is trained on the data originating from X_i, but demonstrates strong performance when confronted with images coming from previously unseen sources, denoted as X_t. This generalizability across unseen sources is a crucial objective of our detector.

Overall Architecture

With the primary objective of bolstering generalizability to unseen sources, we have devised a frequency domain learning network to effectively enhance deepfake detection capabilities. The comprehensive architecture of the FreqNet approach is depicted in Figure 3. Within our methodology, we introduce practical and compact frequency learning plugin modules designed to compel the CNN classifier to operate within the frequency domain. These modules include the high-frequency representation and the frequency convolutional layer. The modular nature of these components allows seamless integration into the CNN classifier. This architectural innovation has culminated in the creation of a novel and lightweight detector, FreqNet, incorporating the frequency learning plugins alongside a limited number of CNN layers.

High-Frequency Representation of Images High-frequency artifacts have been recognized as valuable indicators for distinguishing between real and fake images (Qian et al. 2020; Luo et al. 2021). Consequently, we leverage this valuable insight by adopting the high-frequency components of images as the input for the detector in our approach. In the case of each training image denoted as $x \in \mathbb{R}^{W \times H \times 3}$, our initial step involves converting it into the frequency domain using the Fast Fourier Transform (FFT). Subsequently, we proceed to extract the high-frequency components, represented as f_h, through the application of a high-pass filter denoted as $\mathcal{B}_h$:

$$f_h = \mathcal{B}_h(\mathcal{F}(x)), \tag{3}$$

where $\mathcal{F}$ denotes FFT, $f_h \in \mathbb{R}^{W \times H \times 3}$ denotes the frequency repersentation of images. The zero-frequency is shifted to the center. The high-pass filter $\mathcal{B}_h(\cdot)$ can be defined by:

$$\mathcal{B}_h(f_{i,j}) = \begin{cases} f_{i,j}, otherwise, \\ 0, if \; |i| < W/4, |j| < H/4 \end{cases} \tag{4}$$

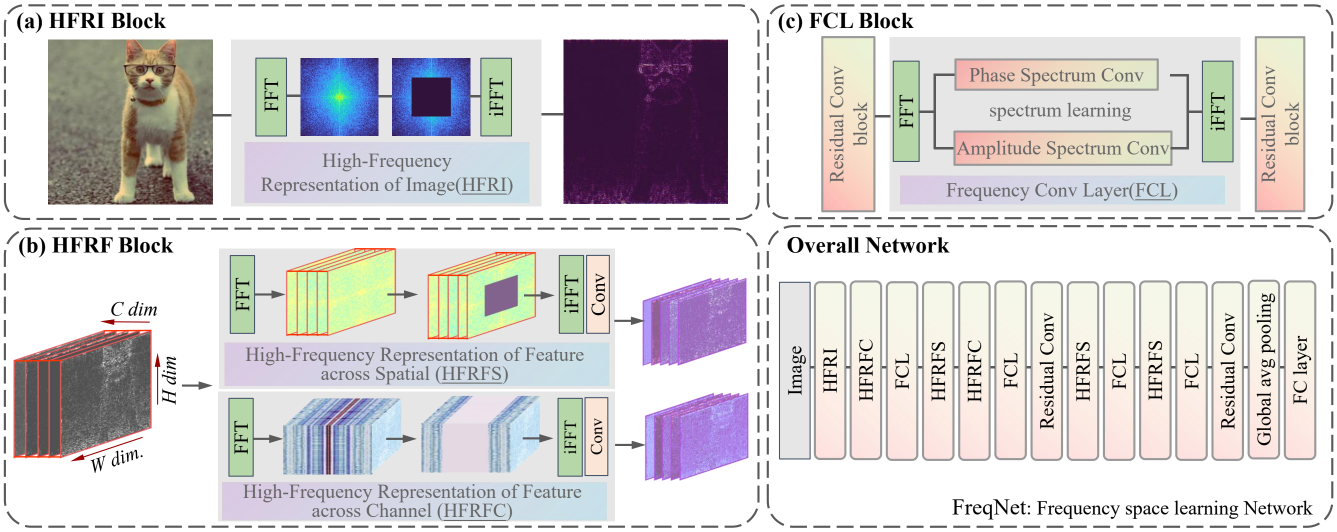


Figure 3: Architecture of FreqNet for generalizable deepfake detection. To augment the capacity for generalization, our FreqNet focuses on the enhancement of frequency spectrum information, prioritizing frequency domain learning within the classifier, consisting of (a) High-Frequency Representation of Image(HFRI), (b) High-Frequency Representation of Feature(HFRF), and (c) Frequency Conv Layer(FCL).

where the center of the image is adopted as the origin.

Following the extraction of high-frequency components, we proceed to transform this frequency information back to image space:

$$x_h = \mathcal{IF}(f_h), \tag{5}$$

where the $\mathcal{IF}$ denotes inverse Fast Fourier Transform (iFFT). The result of this transformation, denoted as x_h, represents the high-frequency components within the image space. This process ensures that our focus remains on the pertinent high-frequency information.

High-Frequency Representation of Feature Indeed, the varying GAN architectures incorporate distinct frequency patterns, which can potentially lead the detector to overfit to the specifics of the training source. To mitigate this issue and enhance the detector's generalization capacity, we adopt a strategy that involves compelling the detector to consistently prioritize and focus on high-frequency information within the feature space. By emphasizing the importance of high-frequency cues in the feature space, we effectively counteract the overfitting tendencies, promoting a more robust and adaptable deepfake detection capability.

Specifically, for output of the k_{th} convolutional layer denoted as $M^k \in \mathbb{R}^{H\times W\times C}$, we transform it to frequency space across spatial (W, H) and the channel dimension C using the Fast Fourier Transform (FFT), respectively. The zero frequency component is moved to the center. Subsequently, we apply a high-pass filter $\mathcal{B}_h$ to extract the high-frequency information from this transformed representation. Finally, we reverse this frequency transformation, converting the extracted high-frequency information back to the feature space:

$$M_h^k(dim) = \begin{cases} \mathcal{IF}_{W,H}(\mathcal{B}_h(\mathcal{F}_{W,H}(M^k))), dim = W, H \\ \quad \mathcal{IF}_C(\mathcal{B}_h(\mathcal{F}_C(M^k))), dim = C \end{cases} \tag{6}$$

where $M_h^k(W, H)$, $M_h^k(C)$ represent the high-frequency components of feature maps M^k acorss spatial (W, H) and channel C dimension in feature space, respectively. We implement two distinct high-frequency component extractors, each targeting different CNN layers within our method. This strategic differentiation allows us to leverage the unique information present at different stages of the network, enhancing the overall sensitivity to high-frequency cues.

Frequency Convolutional Layer Many existing frequency-based approaches follow a paradigm of extracting frequency information from images and employing it to train a CNN classifier. However, this approach often results in the detector overfitting to the specifics of the training source, leading to suboptimal performance when faced with previously unseen sources. In contrast, our approach not only employs the frequency information as the artifact representation. We also introduce frequency space learning as a strategy to significantly enhance the generalization ability of the detector.

Specifically, within our approach, the feature maps from the CNN layers are initially transformed from the feature space to the frequency domain. Following this transformation, we apply convolutional layers on both the phase spectrum and the amplitude spectrum, thereby enabling the detector to learn within the frequency space. Subsequently, the learned spectrum information is transformed back to the feature space using the inverse Fast Fourier Transform. This comprehensive process effectively emphasizes the representation ability of the detector within the frequency domain, enhancing its sensitivity to critical features present in this space.

Let's consider the given feature maps $M^k \in \mathbb{R}^{W \times H \times C}$ from k_{th} CNN layer. The learning process within the frequency space can be formally defined as follows:

$$\begin{aligned} f &= f_{am} + f_{ph}\mathrm{i} = \mathcal{F}_{W,H}(M^k) \\ \widetilde{f_{am}} &= L_{conv}(f_{am}) \\ \widetilde{f_{ph}} &= L_{conv}(f_{ph}) \\ \widetilde{M^k} &= \mathcal{IF}_{W,H}(\widetilde{f_{am}} + \widetilde{f_{ph}}\mathrm{i}) \end{aligned} \tag{7}$$

where $\mathcal{F}_{W,H}, \mathcal{IF}_{W,H}$ denote FFT and iFFT, f_{am}, f_{ph} are amplitude spectrum and phase spectrum of feature maps M^k, and L_{conv} denotes a CNN layer, $\widetilde{f_{am}}, \widetilde{f_{ph}}$ are the learned amplitude spectrum and phase spectrum of feature maps M^k. Subsequently, The feature map $\widetilde{M^k}$ learned in spectrum space can be calculated by iFFT.

As a result, the comprehensive training procedure for FreqNet can be formally defined as:

$$D_{freq} = \arg\min_{\theta} \; l(D_{freq}(x_h; \theta), \; y), \tag{8}$$

where $l()$ denotes the standard cross entropy loss, and D_{freq} is our frequency sapce learning network. Our FreqNet harnesses spectrum learning to accomplish domain-invariant deepfake detection. Within our approach, we have meticulously designed two key modules: the high-frequency representation module and the frequency convolutional layer. Importantly, this work places emphasis on training our detector using a constrained amount of training data, followed by comprehensive evaluations in the challenging wild scenes, encompassing a diverse set of 17 GAN models.

Experiments

In this section, we provide a comprehensive evaluation of the FreqNet. We cover various aspects, including datasets, implementation details, detection performance, and more details to be described. Further elaborations on each of these aspects will be presented to offer a comprehensive understanding of the capabilities and effectiveness of FreqNet.

Datasets

Training set. To ensure a consistent basis for comparison, we employ the training set of ForenSynths (Wang et al. 2020) to train the detectors, aligning with baselines (Wang et al. 2020; Jeong et al. 2022a,c). The training set consists of 20 distinct categories, each comprising 18,000 synthetic images generated using ProGAN, alongside an equal number of real images sourced from the LSUN dataset. In line with previous research (Jeong et al. 2022a,c), we adopt specific 1-class, 2-class, and 4-class training settings, denoted as (horse), (chair, horse), (car, cat, chair, horse), respectively.

Real-world Scene Test set. To assess the generalization ability of the proposed method on the real-world scene, we adopt various images and diverse GAN models. Firstly, we employ the test set of ForenSynths for evaluation. It includes fake images generated by 8 generation model [2]. The real images are sampled from 6 datasets [3]. Additionally, to replicate the unpredictability of wild scenes, we extend our evaluation by collecting images generated by 9 additional GANs [4]. There are 36K test images, with equal numbers of real and fake images. Simultaneously, we curate a dedicated face test set comprising 20,000 real images sourced from Celeba-HQ (Karras et al. 2018), and 60,000 fake face images from ProGAN (Karras et al. 2018), StyleGAN (Karras et al. 2019), and StyleGAN2 (Karras et al. 2020).

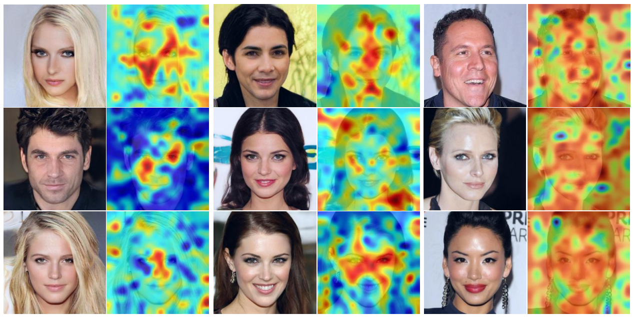

(a) ProGAN (b) StyleGAN (c) CelabA-HQ

Figure 4: The visualization of Class Activate Map (CAM) (Zhou et al. 2016) extracted from detector on face images.

Implementation Details

We design a lightweight CNN classifier, employing residual convolutional blocks without pretraining. During the training process, we utilize the Adam optimizer (Kingma et al. 2015) with an initial learning rate of 2×10^{-2}. The batch size is set at 32, and we train the model for 100 epochs. A learning rate decay strategy is employed, reducing the learning rate by twenty percent after every ten epochs. Consistent with established baselines (Jeong et al. 2022a,c), we utilize the average precision score (A.P.) and accuracy (Acc.) as the primary evaluation metrics to gauge the effectiveness of our proposed method. These metrics provide a comprehensive assessment of the performance of our approach against the baselines. We employ the PyTorch framework (Paszke et al. 2019) for the implementation of our method, utilizing the computational power of the Nvidia GeForce RTX 3090 GPU. For the critical task of Fast Fourier Transform (FFT), we leverage the $torch.fft.fftn$ function within the PyTorch library.

[2] ProGAN (Karras et al. 2018), StyleGAN (Karras et al. 2019), StyleGAN2 (Karras et al. 2020), BigGAN (Brock et al. 2018), CycleGAN (Zhu et al. 2017), StarGAN (Choi et al. 2018), GauGAN (Park et al. 2019) and Deepfake (Rossler et al. 2019)

[3] LSUN (Yu et al. 2015), ImageNet (Russakovsky et al. 2015), CelebA (Liu et al. 2015), CelebA-HQ (Karras et al. 2018), COCO (Lin et al. 2014), and FaceForensics++ (Rossler et al. 2019)

[4] AttGAN(He et al. 2019), BEGAN(Berthelot et al. 2017), CramerGAN(Bellemare et al. 2017), InfoMaxGAN(Lee et al. 2021), MMDGAN(Li et al. 2017), RelGAN(Nie et al. 2019), S3GAN(Lučić et al. 2019), SNGAN(Miyato et al. 2018), and STGAN(Liu et al. 2019)

Methods	Settings		Test Models																	
	Input	#n	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
			Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Wang(2020)	Img	1	50.4	63.8	50.4	79.3	68.2	94.7	50.2	61.3	50.0	52.9	50.0	48.2	50.3	67.6	50.1	51.5	52.5	64.9
Frank(2020)	Freq	1	78.9	77.9	69.4	64.8	67.4	64.0	62.3	58.6	67.4	65.4	60.5	59.5	67.5	69.1	52.4	47.3	65.7	63.3
Durall(2020)	Freq	1	85.1	79.5	59.2	55.2	70.4	63.8	57.0	53.9	66.7	61.4	99.8	99.6	58.7	54.8	53.0	51.9	68.7	65.0
F3Net(2020)	Freq	1	96.9	99.9	86.3	99.8	80.5	99.8	66.6	72.2	76.7	84.0	99.1	100.0	59.1	60.6	61.2	82.3	78.3	87.3
BiHPF(2022a)	Freq	1	82.5	81.4	68.0	62.8	68.8	63.6	67.0	62.5	75.5	74.2	90.1	90.1	73.6	92.1	51.6	49.9	72.1	72.1
FrePGAN(2022c)	Img	1	95.5	99.4	80.6	90.6	77.4	93.0	63.5	60.5	59.4	59.9	99.6	100.0	53.0	49.1	70.4	81.5	74.9	79.3
LGrad (2023)	Grad	1	99.4	99.9	96.0	99.6	93.8	99.4	79.5	88.9	84.7	94.4	99.5	100.0	70.9	81.8	66.7	77.9	86.3	92.7
Ojha (2023)	Fea	1	99.1	100.0	77.2	95.9	69.8	95.8	94.5	99.0	97.1	99.9	98.0	100.0	95.7	100.0	82.4	91.7	89.2	**97.8**
FreqNet	Freq	1	98.0	99.9	92.0	98.7	89.5	97.9	85.5	93.1	96.1	99.1	94.2	98.4	91.8	99.6	69.8	94.4	**89.6**	97.6
Wang(2020)	Img	2	64.6	92.7	52.8	82.8	75.7	96.6	51.6	70.5	58.6	81.5	51.2	74.3	53.6	86.6	50.6	51.5	57.3	79.6
Frank(2020)	Freq	2	85.7	81.3	73.1	68.5	75.0	70.9	76.9	70.8	86.5	80.8	85.0	77.0	67.3	65.3	50.1	55.3	75.0	71.2
Durall(2020)	Freq	2	79.0	73.9	63.6	58.8	67.3	62.1	69.5	62.9	65.4	60.8	99.4	99.4	67.0	63.0	50.5	50.2	70.2	66.4
F3Net(2020)	Freq	2	97.9	100.0	84.5	99.5	82.2	99.8	65.5	73.4	81.2	89.7	100.0	100.0	57.0	59.2	59.9	83.0	78.5	88.1
BiHPF(2022a)	Freq	2	87.4	87.4	71.6	74.1	77.0	81.1	82.6	80.6	86.0	86.6	93.8	80.8	75.3	88.2	53.7	54.0	78.4	79.1
FrePGAN(2022c)	Img	2	99.0	99.9	80.8	92.0	72.2	94.0	66.0	61.8	69.1	70.3	98.5	100.0	53.1	51.0	62.2	80.6	75.1	81.2
LGrad (2023)	Grad	2	99.8	100.0	94.8	99.7	92.4	99.6	82.5	92.4	85.9	94.7	99.7	99.9	73.7	83.2	60.6	67.8	86.2	92.2
Ojha (2023)	Fea	2	99.7	100.0	78.8	97.4	75.4	96.7	91.2	99.0	91.9	99.8	96.3	99.9	91.9	100.0	80.0	89.4	88.1	97.8
FreqNet	Freq	2	99.6	100.0	90.4	98.9	85.8	98.1	89.0	96.0	96.7	99.8	97.5	100.0	88.0	98.8	80.7	92.0	**91.0**	**97.9**
Wang(2020)	Img	4	91.4	99.4	63.8	91.4	76.4	97.5	52.9	73.3	72.7	88.6	63.8	90.8	63.9	92.2	51.7	62.3	67.1	86.9
High-Freq	Freq	4	98.9	100.0	74.4	98.3	68.8	97.3	75.2	92.1	71.0	87.9	92.7	100.0	75.5	86.5	57.0	74.9	76.7	92.1
Frank(2020)	Freq	4	90.3	85.2	74.5	72.0	73.1	71.4	88.7	86.0	75.5	71.2	99.5	99.5	69.2	77.4	60.7	49.1	78.9	76.5
Durall(2020)	Freq	4	81.1	74.4	54.4	52.6	66.8	62.0	60.1	56.3	69.0	64.0	98.1	98.1	61.9	57.4	50.2	50.0	67.7	64.4
F3Net(2020)	Freq	4	99.4	100.0	92.6	99.7	88.0	99.8	65.3	69.9	76.4	84.3	100.0	100.0	58.1	56.7	63.5	78.8	80.4	86.2
BiHPF(2022a)	Freq	4	90.7	86.2	76.9	75.1	76.2	74.7	84.9	81.7	81.9	78.9	94.4	94.4	69.5	78.1	54.4	54.6	78.6	77.9
FrePGAN(2022c)	Img	4	99.0	99.9	80.7	89.6	84.1	98.6	69.2	71.1	71.1	74.4	99.9	100.0	60.3	71.7	70.9	91.9	79.4	87.2
LGrad (2023)	Grad	4	99.9	100.0	94.8	99.9	96.0	99.9	82.9	90.7	85.3	94.0	99.6	100.0	72.4	79.3	58.0	67.9	86.1	91.5
Ojha (2023)	Fea	4	99.7	100.0	89.0	98.7	83.9	98.4	90.5	99.1	87.9	99.8	91.4	100.0	89.9	100.0	80.2	90.2	89.1	98.3
FreqNet	Freq	4	99.6	100.0	90.2	99.7	88.0	99.5	90.5	96.0	95.8	99.6	85.7	99.8	93.4	98.6	88.9	94.4	**91.5**	**98.5**

Table 1: Cross-model performance on the test set of ForenSynths(Wang et al. 2020). Bold and underline represent the best and second-best performance, respectively.

Method	AttGAN		BEGAN		CramerGAN		InfoMaxGAN		MMDGAN		RelGAN		S3GAN		SNGAN		STGAN		Mean	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Wang(2020)	51.1	83.7	50.2	44.9	81.5	97.5	71.1	94.7	72.9	94.4	53.3	82.1	55.2	66.1	62.7	90.4	63.0	92.7	62.3	82.9
F3Net(2020)	85.2	94.8	87.1	97.5	89.5	99.8	67.1	83.1	73.7	99.6	98.8	100.0	65.4	70.0	51.6	93.6	60.3	99.9	75.4	93.1
LGrad (2023)	68.6	93.8	69.9	89.2	50.3	54.0	71.1	82.0	57.5	67.3	89.1	99.1	78.5	86.0	78.0	87.4	54.8	68.0	68.6	80.8
Ojha (2023)	78.5	98.3	72.0	98.9	77.6	99.8	77.6	98.9	77.6	99.7	78.2	98.7	85.2	98.1	77.6	98.7	74.2	97.8	77.6	**98.8**
FreqNet	89.8	98.8	98.8	100.0	95.2	98.2	94.5	97.3	95.2	98.2	100.0	100.0	88.3	94.3	85.4	90.5	98.8	100.0	**94.0**	97.5

Table 2: Cross-model performance on the self-synthesis dataset.

Methods	Parameters ↓	mAcc. ↑ of 17 models
F3Net(2020)	48.9 M	77.8
LGrad(2023)	46.6 M	76.8
Ojha(2023)	304.0 M	83.0
FreqNet	**1.9 M**	**92.8(+9.8)**

Table 3: Comparison of Parameters.

Deepfake Performance on Real-world Scene

In order to demonstrate the remarkable generalization ability of our FreqNet on unseen sources, we carry out evaluations on a real-world scene dataset. This dataset comprises images sourced from a total of 17 different generation models, encompassing 8 models from the ForenSynths test set and an additional 9 models from our own self-synthesis process. This evaluation setup introduces increased complexity compared to the previous experiments, as the testing sets encompass a diverse range of GANs. This challenging scenario effectively simulates an open-world scene, making it a rigorous test of our approach's adaptability and robustness in real-world, unpredictable settings.

We compare with the previous methods: BiHPF (Jeong et al. 2022a), FreGAN (Jeong et al. 2022c), LGrad (Tan et al.

2023), Ojha (Ojha et al. 2023). To ensure fair and meaningful comparisons, we adopt the same experimental setting as the established baselines (Jeong et al. 2022a,c). Specifically, the 1-class, 2-class and 4-class settings refer to training with the *(horse)*, *(chair, horse)*, *(car, cat, chair, horse)* categories of ProGAN, respectively.

The results of the ForenSynths dataset are presented in Table 1. The proposed FreqNet surpasses its counterparts in terms of mean Acc. metric and mean A.P. metrics, except for the value of A.P. on the 1-class setting. Notably, with the 4-class setting, FreqNet achieves a mean Acc. value of 91.5%, demonstrating its strong performance. In comparison to the current state-of-the-art methods LGrad and Ojha, our FreqNet exhibits substantial improvements, surpassing these methods by 5.4% and 2.4% in mean Acc., using fewer parameters. In the 1- and 2-class settings, our FreqNet also achieves gains of 2.9% and 0.4% compared to Ojha. Furthermore, compared to FingerprintNet(Jeong et al. 2022b) tested on six unseen models, our FreqNet achieves a marked improvement in mean accuracy, rising from 82.6% to 90.6% and exhibiting a significant gain of 8.0%. Additionally, we provide results on 9 models from self-synthesis in Table 2. We adopt the 4-class setting detector to perform testing. Compared to Ojha, our FreqNet achieves a marked improvement in mean accuracy, soaring from 77.6% to an impressive 94.0%, resulting in a significant gain of 16.4%. When testing on face images, the proposed FreqNet achieves accuracy rates of 98.7%, 99.0%, and 99.5% on the ProGAN, StyleGAN, and StyleGAN2 datasets, respectively.

Furthermore, we provide a comprehensive overview of the number of parameters and the mean accuracy across all real-world scenes in Table 3. It is evident that our FreqNet, with a modest parameter count of 1.9 million, significantly outperforms the current state-of-the-art model Ojha (Ojha et al. 2023), which boasts an extensive 304 million parameters. This substantial difference in parameter count translates into a notable performance gain, with our FreqNet achieving a remarkable improvement of 9.8% in mean accuracy compared to the larger model. This result underscores the efficiency and effectiveness of our FreqNet approach, demonstrating that superior performance can be achieved with significantly fewer parameters, a crucial advantage in real-world applications.

Compared to other frequency-based methods, such as BiHPF (Jeong et al. 2022a), FrePGAN (Jeong et al. 2022c), F3Net(Qian et al. 2020), our FreqNet achieves better performance on the real-world scene. The results confirm the generalization capability of the proposed frequency domain learning to extract a general representation of artifacts, and generalize this representation across various GAN models and categories.

We perform ablation analyses on our FreqNet by individually removing the proposed modules. The results of these ablation experiments are presented in Table 4. Upon removal of the designed modules, we observe a decline in the detection performance, underscoring the efficacy of the proposed components. In the revised version, we will expound further on the specifics of the ablation analysis to provide a more comprehensive understanding.

HFRI	HFRFS	HFRFC	FCL	mean Acc.
	✓	✓	✓	84.3
✓		✓	✓	85.3
✓	✓		✓	87.8
✓			✓	82.0
✓	✓	✓		83.8
✓	✓	✓	✓	91.5

Table 4: Ablation Study of FreqNet on the ForenSynths(Wang et al. 2020).

Visualization of Class Activate Map. To visually demonstrate the discriminative regions identified by our detector, we present the Class Activation Maps (CAM) in Figure 4. The CAMs are generated using images from ProGAN, StyleGAN, StyleGAN2, and CelebA-HQ. The CAMs provide insights into the areas of focus for the detector in distinguishing real from fake images. It's noteworthy that the CAMs for real images highlight a broader portion of the image, while the CAMs for fake images tend to emphasize localized regions. Interestingly, even though the detector is primarily trained using a dataset containing cars, cats, chairs, and horses, it showcases the ability to recognize face images effectively. This highlights the versatility and adaptability of our detector in identifying distinct deepfake characteristics, even beyond the classes it was primarily trained on.

Conclusion

This study has focused on the introduction of FreqNet, a lightweight frequency space learning network designed for the task of generalizable forgery image detection. Our approach capitalizes on the power of frequency domain learning, offering an adaptable solution for the challenging problem of deepfake detection across diverse sources and GAN models. Within our methodology, we introduce practical and compact frequency learning plugin modules designed to compel the CNN classifier to operate within the frequency domain. The extensive experiments conducted on 17 different generation models serve as compelling evidence of FreqNet's generalization ability. This research contributes to advancing the field of deepfake detection, showcasing the potential of FreqNet to effectively combat the challenges posed by evolving forgery techniques and diverse image sources.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No.2021ZD0112100), National NSF of China (No.U1936212, No.62120106009), National Natural Science Foundation of China under Grants 62072394, Natural Science Foundation of Hebei province under Grant F2021203019, A*STAR Career Development Funding (CDF) Award (Grant No:222D800031) and the Fundamental Research Funds for the Central Universities+2022YJS026.

References

Bellemare, M. G.; et al. 2017. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*.

Berthelot, D.; et al. 2017. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*.

Brock, A.; et al. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Cao, J.; et al. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4122.

Chai, L.; et al. 2020. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, 103–120. Springer.

Chen, L.; et al. 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18710–18719.

Choi, Y.; et al. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.

Cooley, J. W.; et al. 1969. The fast Fourier transform and its applications. *IEEE Transactions on Education*, 12(1): 27–34.

Durall, R.; et al. 2020. 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*, 7890–7899.

Frank, J.; et al. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, 3247–3258. PMLR.

Goodfellow, I. J.; et al. 2014. Generative Adversarial Nets. In *NIPS*.

Haliassos, A.; et al. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5039–5049.

He, Y.; et al. 2021. Beyond the Spectrum: Detecting Deepfakes via Re-Synthesis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2534–2541. International Joint Conferences on Artificial Intelligence Organization.

He, Z.; et al. 2019. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, 28(11): 5464–5478.

Jeong, Y.; et al. 2022a. BiHPF: Bilateral High-Pass Filters for Robust Deepfake Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 48–57.

Jeong, Y.; et al. 2022b. FingerprintNet: Synthesized Fingerprints for Generated Image Detection. In *European Conference on Computer Vision*, 76–94. Springer.

Jeong, Y.; et al. 2022c. FrePGAN: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1060–1068.

Karras, T.; et al. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Karras, T.; et al. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Karras, T.; et al. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kingma, D. P.; et al. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Lee, K. S.; et al. 2021. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3942–3952.

Li, C.; Huang, Z.; Paudel, D. P.; Wang, Y.; Shahbazi, M.; Hong, X.; and Van Gool, L. 2023. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1339–1349.

Li, C.-L.; et al. 2017. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30.

Li, J.; et al. 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6458–6467.

Li, Y.; et al. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, 1–7. IEEE.

Lin, T.-Y.; et al. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, M.; et al. 2019. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3673–3682.

Liu, Z.; et al. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Lučić, M.; et al. 2019. High-fidelity image generation with fewer labels. In *International conference on machine learning*, 4183–4192. PMLR.

Luo, Y.; et al. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16317–16326.

Masi, I.; et al. 2020. Two-branch recurrent network for isolating deepfakes in videos. In *European conference on computer vision*, 667–684. Springer.

Miyato, T.; et al. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Nie, W.; et al. 2019. Relgan: Relational generative adversarial networks for text generation. In *International conference on learning representations*.

Ojha, U.; et al. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.

Park, T.; et al. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346.

Paszke, A.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qian, Y.; et al. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, 86–103. Springer.

Rossler, A.; et al. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Russakovsky, O.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Shiohara, K.; et al. 2022. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720–18729.

Tan, C.; et al. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12105–12114.

Wang, C.; et al. 2021. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14923–14932.

Wang, S.-Y.; et al. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695–8704.

Wang, Y.; et al. 2023. Dynamic Graph Learning With Content-Guided Spatial-Frequency Relation Reasoning for Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7278–7287.

Woo, S.; et al. 2022. ADD: Frequency Attention and Multi-View Based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 122–130.

Yu, F.; et al. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Yu, Y.; et al. 2020. Mining generalized features for detecting ai-manipulated fake faces. *arXiv preprint arXiv:2010.14129*.

Zhou, B.; et al. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhu, J.-Y.; et al. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.