# Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues

**Abstract**

As realistic facial manipulation technologies have achieved remarkable progress, social concerns about potential malicious abuse of these technologies bring out an emerging research topic of face forgery detection. However, it is extremely challenging since recent advances are able to forge faces beyond the perception ability of human eyes, especially in compressed images and videos. We find that mining forgery patterns with the awareness of frequency could be a cure, as frequency provides a complementary viewpoint where either subtle forgery artifacts or com- pression errors could be well described. To introduce frequency into the face forgery detection, we propose a novel Frequency in Face Forgery Network (F3-Net), taking advantages of two different but complemen- tary frequency-aware clues, frequency-aware decomposed image com- ponents, and local frequency statistics, to deeply mine the forgery patterns via our two-stream collaborative learning framework. We apply DCT as the applied frequency-domain transformation. Through compre- hensive studies, we show that the proposed F3-Net significantly outper- forms competing state-of-the-art methods on all compression qualities in the challenging FaceForensics++ dataset, especially wins a big lead upon low-quality media.

**Keywords**：Face Forgery Detection, Frequency, Collaborative Learning

## 1 Introduction

Frequency-Based Forgery Detection. Frequency domain analysis is a classical and important method in image signal processing and has been widely used in a number of applications such as image classification , steganalysis , texture classification and super-resolution.Recently, several attempts have been made to solve forgery detection using frequency cues. Some studies use Wavelet Transform (WT) or Discrete Fourier Transform (DFT) to convert pictures to frequency domain and mine underlying artifacts.

### 1.1 Traditional method

Conventional frequency domains, such as FFT and DCT, do not match the shift-invariance and local consistency owned by nature images, thus vanilla CNN structures might be infeasible. As a result, CNN-compatible frequency representation becomes pivotal if we would like to leverage the dis- criminative representation power of learnable CNNs for frequency-aware face forgery detection.

### 1.2 Mining frequency-aware clues

We would like to introduce two frequency-aware forgery clues that are compatible with the knowledge mining by deep convolutional networks. From one aspect, it is possible to decompose an image by separating its frequency signals, while each decomposed image component indicates a certain band of frequencies. The first frequency-aware forgery clue is thus discovered by the intuition that we are able to identify subtle forgery artifacts that are somewhat salient (i.e., in the form of unusual patterns) in the decomposed com- ponents with

higher frequencies, as the examples shown in the middle column of Fig. 1. This clue is compatible with CNN structures, and is surprisingly robust to compression artifacts. From the other aspect, the decomposed image components describe the frequency-aware patterns in the spatial domain, but not explicitly render the frequency information directly in the neural networks. We suggest the second frequency-aware forgery clue as the local frequency statistics. In each densely but regularly sampled local spatial patch, the statistics is gathered by counting the mean frequency responses at each frequency band. These frequency statistics reassemble back to a multichannel spatial map, where the number of channels is identical to the number of frequency bands.

## 2 Method

### 2.1 Overview

Overview of the F3-Net. The proposed architecture consists of three novel methods: FAD for learning subtle manipulation patterns through frequency-aware im- age decomposition; LFS for extracting local frequency statistics and MixBlock for collaborative feature interaction. Figure1：

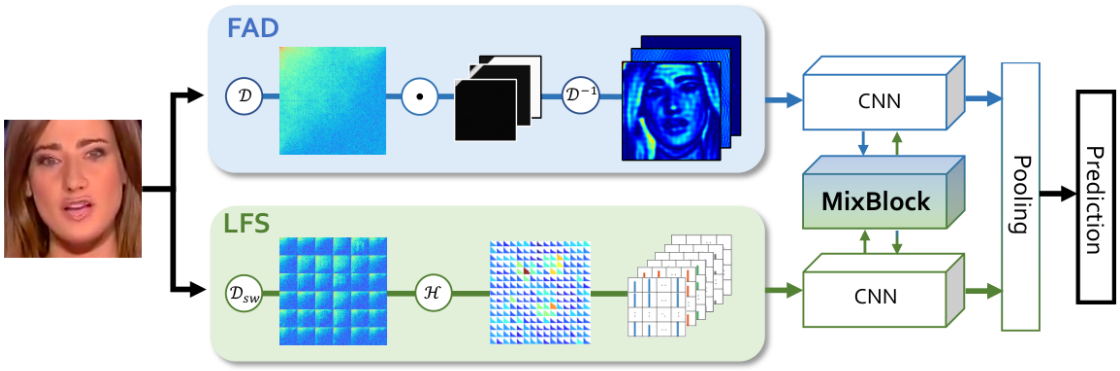


Figure 1: Overview of the method

Fig. 2. (a) The proposed Local Frequency Statistics (LFS) to extract local frequency domain statistical information. SWDCT indicates applying Sliding Window Discrete Cosine Transform and H indicates gathering statistics on each grid adaptively. (b) Extracting statistics from a DCT power spectrum graph, $\oplus$ indicates element-wise addition and $a \odot b$ indicates element-wise multiplication.

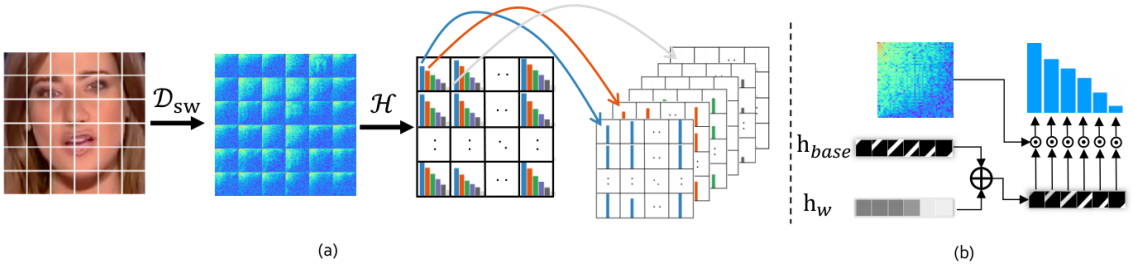


Figure 2: Overview of the method LFS

Fig. 3. (a) The proposed Frequency-aware Decomposition (FAD) to discover salient frequency components. D indicates applying Discrete Cosine Transform (DCT). $D^{-1}$ indicates applying Inversed Discrete Cosine Transform (IDCT). Several frequency band components can be concatenated together to extract a wider range of information. (b) The distribution of the DCT power spectrum. We flatten 2D power spectrum to 1D

by summing up the amplitudes of each frequency band. We divide the spectrum into 3 bands with roughly equal energy.. Figure3：

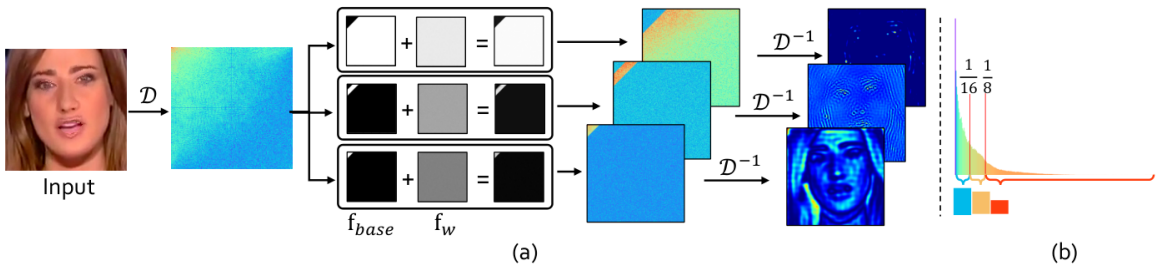


Figure 3: Overview of the method FAD

### 2.2 Loss

cross-entropy loss:

$$Loss = \frac{1}{batch\_size} \sum_{j=1}^{batch\_size} \sum_{i=1}^{n} [-y_{ji} \log \widehat{y_{ji}} - (1 - y_{ji}) \log(1 - \widehat{y_{ji}})]$$

## 3 Implementation details

### 3.1 improved method

After processing the images in the original architecture with FAD and LFS, We first tried to replace xception with the more powerful inceptionV4 architecture, but found that the effect was not good. In addition,we adopt a new method of extracting the corresponding amplitude and phase of the image after Fourier transform, and then inputting it into the subsequent network. The reason why we perform Fourier transform on images is because the original images are processed in the spatial domain, and the extracted features are weak. After being converted to the frequency space, information from different frequency bands gathers together, such as high-frequency and high-frequency, and low-frequency and low-frequency together, making the feature information more compact and dense.

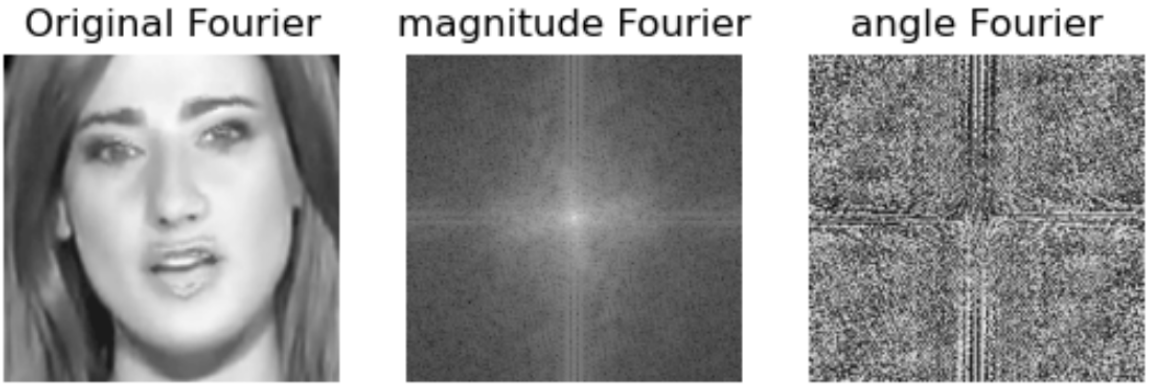


Figure 4: Overview of the innovate method

Typically, this operation is independently conducted over the spatial dimension of each individual channel. Given an image $x \in R^{H \times W \times C}$, the Fourier transform F( • ) converts it to Fourier space as the complex

component F(x), which is expressed as:

$$F(x)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{N-1} x(\mathrm{h},w) e^{j(-2\pi)(\frac{h}{H}u + \frac{\mathrm{w}}{W}v)} \tag{1}$$

From Euler's formula, it can be inferred that:

$$e^{j(-2\pi)(\frac{h}{h}u + \frac{\mathrm{w}}{W}v)} = \cos\{(-2\pi)(\frac{h}{H}u + \frac{\mathrm{w}}{W}v)\} + j\sin\{(-2\pi)(\frac{h}{H}u + \frac{\mathrm{w}}{W}v)\} \tag{2}$$

The real and imaginary parts of the above formula are R(x) and I(x):

$$R(X)(u,v) = \cos\{(-2\pi)(\frac{h}{H}u + \frac{\mathrm{w}}{W}v)\} \tag{3}$$

$$I(x)(u,v) = \sin\{(-2\pi)(\frac{h}{H}u + \frac{\mathrm{w}}{W}v)\} \tag{4}$$

The amplitude component A(x)(u, v) and phase component P(x)(u, v) are expressed as:

$$A(x)(u,v) = \sqrt{R^2(x)(u,v) + I^2(x)(u,v)} \tag{5}$$

$$P(x)(u,v) = \arctan[\frac{I(x)(u,v)}{R(x)(u,v)}] \tag{6}$$

### 3.2 experimental design

Because the FaceForensics++ data set is so large,(FaceForensics++:The original downladed source videos from youtube.All h264 compressed videos with compression rate factor.All raw extracted images as pngs: 2TB),the college's AiStation cluster only allocates 200GB of storage space to each account, and I had to replace it with a smaller data set.With a small data set, use the original method of the paper for training and improve the original method. To study the transferability of classifiers trained to detect CNN-generated images, we collected a dataset of images created from a variety of CNN models.Our dataset contains 11 synthesis models.We chose methods that span a variety of CNN architectures, datasets,and losses.All of these models have an upsampling convolutional structure since this is by far the most common design for generative CNNs.

There are two options here: if your GPU resources are sufficient, you can use the training set data to retrain the weight parameters. If you do not have enough GPU resources, you can use the pretrained xception and Inception v4 architecture. The weights of the pretrained model are trained based on the imageNet dataset. Our experiment here is based on the weight of restarting training. If you want to use the pretrained model, you need to perform the following processing. In the original paper, the extracted frequencies were divided into four frequency bands, so for the first convolutional kernel (conv1) input of the pretrained xception. The channels dimension has been changed from 3 to 12 dimensions, while the out channels remain unchanged and remain at 32 dimensions. At the same time, the weight of conv1 is divided by 4. The code is as follows:

**Procedure 1** Spatial-Temporal Texture Transformer Network for Video Inpainting.

**Input:** the first convolutional kernel input of the pretrained xception *conv1,* the first convolutional kernel input of the pretrained Inception v4 *features.0.conv,* the number of extracted frequencies were divided into frequency bands *s*

**Output:** complection *weight*

**for** *i* ***in*** *range(s)* **do**

$FADincep.features[0].conv.weight.data[B, i*3:(i+1)*3, H, W]$ = features.0.conv.weight/s

$FAD.xcep.conv1.weight.data[B, i*3:(i+1)*3, H, W]$ = conv1.data/s

**end**

### 3.3 Main contributions

The original article analyzes the difference between real and fake pictures from the perspective of frequency domain, and proposes a combination of FAD (Frequency-Aware Decomposition) and LFS (Local Frequency Statistics) to mine the characteristics of pictures. Randomly divide the range of frequency bands. In the improved method, more attention is paid to the high-frequency part of the image (the part that is difficult to generate by the AIGC model) to find flaws in the AIGC generation model. In addition, it also digs from the perspective of amplitude and phase to find the essential difference between real images and fake images.

## 4 Results and analysis

| test method | original method | | improved method | |
|---|---|---|---|---|
| | accuracy | avg precision | accuracy | avg precision |
| progan | 0.86 | 0.961027345 | 0.9275 | 0.987521137 |
| stylegan | 0.66 | 0.853625538 | 0.906666667 | 0.976471236 |
| biggan | 0.515 | 0.559306129 | 0.545 | 0.659124017 |
| cyclegan | 0.568333333 | 0.637259996 | 0.6625 | 0.81720866 |
| stargan | 0.785 | 0.925476219 | 0.995 | 1 |
| gaugan | 0.49 | 0.490348628 | 0.62 | 0.667698583 |
| crn | 0.535 | 0.893337156 | 0.76 | 0.842724305 |
| imle | 0.535 | 0.741933873 | 0.665 | 0.738450147 |
| deepfake | 0.55 | 0.651522981 | 0.72 | 0.768439957 |
| stylegan2 | 0.66375 | 0.884303506 | 0.86 | 0.941590191 |

The detection accuracy comparison between improved approach and original baselines. The first digit represents accuracy, and the last digit represents average precision.

## 5 Conclusion and future work

The traditional approach of directly inputting image spatial domain information into the network architecture makes it difficult to deeply mine fingerprints. By transforming images into frequency space through Fourier transform for processing, the distribution of information is more orderly, making it easier to mine the hidden information behind artistic images due to network architecture and machine processing methods. The original text only divides different frequencies into sub bands, while the newly proposed method processes them from both amplitude and phase dimensions.