Learnable Frequency Decomposition for Image Forgery Detection and Localization

Dong Li, **Jiaying Zhu**, **Yidi Liu**, **Xin Lu**, **Xueyang Fu***, **Jiawei Liu**, **Aiping Liu**, **Zheng-Jun Zha**
University of Science and Technology of China
{dongli6, zhujy53, liuyidi2023, luxion}@mail.ustc.edu.cn,
{xyfu, jwliu6, aipingl, zhazj}@ustc.edu.cn

Abstract

Concern for image authenticity spurs research in image forgery detection and localization (IFDL). Most deep learning-based methods focus primarily on spatial domain modeling and have not fully explored frequency domain strategies. In this paper, we observe and analyze the frequency characteristic changes caused by image tampering. Observations indicate that manipulation traces are especially prominent in phase components and span both low and high-frequency bands. Based on these findings, we propose a forensic frequency decomposition network (F2D-Net), which incorporates deep Fourier transforms and leverages both phase information and high and low-frequency components to enhance IFDL. Specifically, F2D-Net consists of the Spectral Decomposition Subnetwork (SDSN) and the Frequency Separation Subnetwork (FSSN). The former decomposes the image into amplitude and phase, focusing on learning the semantic content in the phase spectrum to identify forged objects, thus improving forgery detection accuracy. The latter further adaptively decomposes the output of the SDSN to obtain corresponding high and low frequencies, and applies a divide-and-conquer strategy to refine each frequency band, mitigating the optimization difficulties caused by coupled forgery traces across different frequencies, thereby better capturing the pixels belonging to the forged object to improve localization accuracy. Experiments on multiple datasets demonstrate that our method outperforms state-of-the-art image forgery detection and localization techniques both qualitatively and quantitatively.

1 Introduction

With the development of image editing and generation technologies[Goodfellow *et al.*, 2020; Ho *et al.*, 2020; Ramesh *et al.*, 2022], image forgery has become increasingly easier and more prevalent. This advancement has raised widespread public concern, as forgeries pose threats to personal and societal security and privacy, potentially even causing panic. For instance, malicious users can easily modify objects in images using advanced forgery techniques, creating fake news or falsifying evidence in court. Therefore, it is crucial to develop effective and robust methods for detecting and localizing image forgeries.

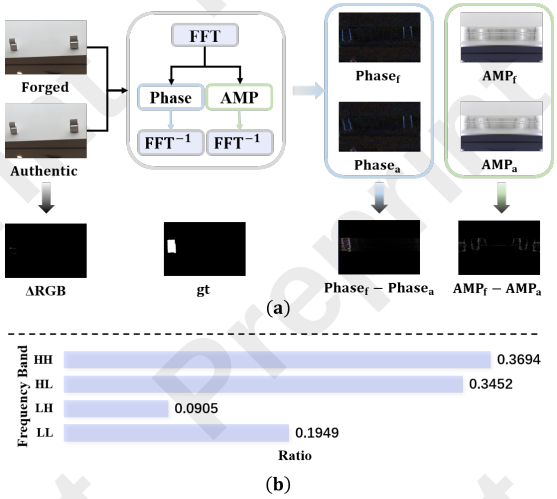

Figure 1: Observing frequency characteristic changes caused by image tampering. (a) The phase and amplitude spectra of real and forged images are subtracted to reveal the changes introduced by tampering. (b) The residual distributions between real and forged images are statistically analyzed across different frequency bands.

In fact, significant progress has been made in forensic forgery detection technologies in recent years. Some works are based on clearly defined low-level features, such as JPEG compression, demosaicking, or interpolation [Bammey *et al.*, 2020]. Certain detectors show favorable results for specific types of image tampering, such as splicing. The academic community has also made milestone achievements in general image forgery detection. For example, RGB-N [Zhou *et al.*, 2018] distinguishes real and manipulated regions by analyzing noise inconsistencies in steganalysis model filters. MVSS-Net [Chen *et al.*, 2021b] learns multi-view features by utilizing noise patterns and edge artifacts. ObjectFormer [Wang *et al.*, 2022a] uses high-frequency information to de-

*Corresponding author.

tect subtle manipulations. However, most existing methods either focus solely on spatial domain forgery features or treat frequency features as a separate modality, with limited exploration of the frequency-domain characteristics presented by forged objects. While artifacts introduced by image editing are not visible in the RGB domain, they often become evident in the frequency domain [Chen *et al.*, 2021a], inspiring us to further investigate the frequency characteristic changes caused by image tampering.

As shown in Figure 1, we apply Fourier transform to both real and forged images to analyze the changes caused by image tampering. On one hand, we subtract their amplitude and phase spectra to observe the changes. To facilitate visualization, the frequency domain residuals, amplitude, and phase spectra in Figure 1a are subjected to inverse Fourier transform. It can be observed that the changes in the image phase components are more closely related to the tampering mask, indicating that **manipulation traces are more easily detectable in the phase component**. This is because most forgery methods typically manipulate images at the object level (semantic content), and the phase components in the Fourier space correspond to the semantic information of the image, while amplitude components correspond to style information [Xu *et al.*, 2021]. On the other hand, we also analyze the impact of forgery on the high- and low-frequency bands of the image. We randomly select a thousand pairs of real and forged images and statistically analyze the residual distributions of real and forged images across different frequency bands. As shown in Figure 1b, **the changes caused by forgery exist in both high and low frequencies**, which is more universal than previous studies that suggest forgery is concentrated in high frequencies [Wang *et al.*, 2022a].

Based on these observations, we propose a learnable forensic frequency decomposition network (F2D-Net) for image forgery detection and localization. Specifically, F2D-Net consists of two components: the Spectrum Decomposition Subnetwork (SDSN) and the Frequency Separation Subnetwork (FSSN). The SDSN decomposes the image into amplitude and phase spectra, focusing on learning the phase spectrum to capture forgery traces and identify forged objects, thereby improving forgery detection accuracy. The SDSN also includes spatial-frequency interaction blocks to facilitate the interaction of forgery features in both spatial and frequency domains, further enhancing detection accuracy. The FSSN is responsible for adaptively decomposing the output of the SDSN into high and low frequencies, and applying a divide-and-conquer strategy to refine each frequency band. This approach mitigates the optimization difficulty caused by the coupling of forgery traces across different frequencies, allowing for better capture of the pixels belonging to the forged object, and thus improving localization accuracy. Our method improves detection and localization accuracy by leveraging phase semantic information and decoupled high- and low-frequency domain information to assist the model in distinguishing and capturing forged objects. In summary, our contributions are as follows:

- We analyze the frequency characteristics changes caused by image tampering and propose a new Fourier-based method for image forgery detection and localization—F2D-Net.
- We develop the spectral decomposition subnetwork, which focuses on learning manipulation traces in the phase spectrum, thereby accurately capturing subtle changes in the phase of forged objects and enhancing forgery detection.
- We design the frequency separation subnetwork to reduce the interference between forgery traces in different frequency bands, further improving the precision of forgery localization.

We conduct extensive experiments on multiple benchmarks and demonstrate that our method outperforms state-of-the-art methods both qualitatively and quantitatively.

2 Related Works

Image forgery detection and localization. Most early works propose to localize a specific type of forgery, including splicing [Huh *et al.*, 2018], copy-move [Cozzolino *et al.*, 2015], and removal [Aloraini *et al.*, 2020]. Although these methods perform well in detecting the specific forgery type, they are obviously insufficient in dealing with real-world cases due to the unknown real forgery types. Therefore, tackling multiple forgery types in one model has been emphasized in recent work. RGB-N [Zhou *et al.*, 2018] uses a dual-stream network to extract RGB and noise features for detecting and localizing image forgery by capturing visual artifacts and modeling region inconsistencies. ManTra-net [Wu *et al.*, 2019] leverages an end-to-end network, which extracts image manipulation trace features and identifies anomalous regions by assessing how different a local feature is from its reference features. SPAN [Hu *et al.*, 2020] attempts to model the spatial correlation via local self-attention blocks and pyramid propagation. MVSS-Net [Chen *et al.*, 2021b] has designed an edge-supervised branch that uses edge residual blocks to capture fine-grained boundary detail in a shallow to deep manner. PSCCNet [Liu *et al.*, 2022] uses a progressive spatial-channel correlation module that uses features at different scales and dense cross-connections to generate operational masks in a coarse-to-fine fashion. HiFi_IFDL [Guo *et al.*, 2023] employs a hierarchical fine-grained approach for IFDL representation learning, utilizing level-wise classification and dependencies for improved performance. In this work, We design a deep Fourier-based network to enhance phase features and adaptively learn low-high frequencies for improved IFDL.

3 Methodology

3.1 Fourier transform of images

For the forgery, some subtle manipulation traces are no longer visible in the spatial domain. Previous works rarely learn the features of tampering artifacts in the frequency domain. To this end, we revisit the forged images via Fourier transform and design a forensic frequency decomposition network to capture the frequency representations of forgery effectively.

As recognized, the Fourier transform is widely used to analyze the frequency content of images. Given a single channel image x with the shape of $H \times W$, the Fourier transform $\mathcal{F}$ converts to the Fourier space as a complex component $\mathcal{F}(x)$,

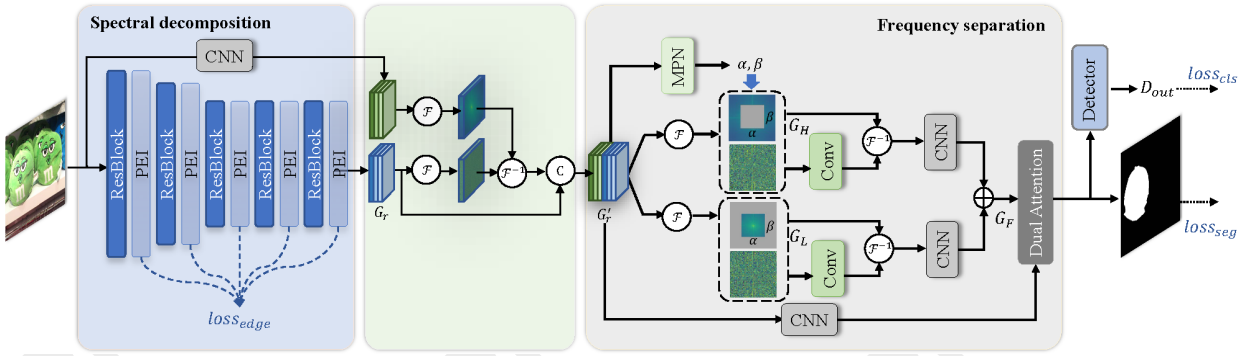

Figure 2: An overview of the proposed framework. The input is a suspicious image ($H \times W \times 3$), and the output is a predicted mask ($H \times W \times 1$), which localizes the forged regions

which is expressed as:

$$\mathcal{F}(x)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w)\, e^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}, \tag{1}$$

and $\mathcal{F}^{-1}(x)$ defines the inverse Fourier transform accordingly. Since an image or feature may contain multiple channels, we separately apply Fourier transform to each channel in our work with the FFT [Frigo and Johnson, 1998]. The amplitude component $\mathcal{A}(x)(u,v)$ and phase component $\mathcal{P}(x)(u,v)$ are expressed as:

$$\begin{aligned} \mathcal{A}(x)(u,v) &= \sqrt{R^2(x)(u,v) + I^2(x)(u,v)}, \\ \mathcal{P}(x)(u,v) &= \arctan\left[\frac{I(x)(u,v)}{R(x)(u,v)}\right], \end{aligned} \tag{2}$$

where $R(x)$ and $I(x)$ represent the real and imaginary parts of $\mathcal{F}(x)$ respectively.

Targeting at Image forgery detection and localization, we employ Fourier transform to conduct the detailed frequency analysis of forgery, as shown in Figure 1. There are two observations in frequency domain: 1) Manipulation traces are more easily detected on the phase components; 2) variations caused by image manipulation are spread over both high and low frequency components. Therefore, it is logical to perform effective learning on the phase component and model low-high frequency components adaptively.

3.2 Overview

Based on the above analysis, we design a simple but effective F2D-Net framework as shown in Figure 2. The input image is represented as $X \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width of the image. First, the image is input to the spectral decomposition subnetwork, obtaining $G_r \in \mathbb{R}^{H_s\times W_s\times C_s}$. Next, G_r and $X_d \in \mathbb{R}^{H_s\times W_s\times C_s}$ obtained by downsampling X are transformed into frequency domain by Fourier transform:

$$\mathcal{A}(G_r), \mathcal{P}(G_r) = \mathcal{F}(G_r), \tag{3}$$

$$\mathcal{A}(X_d), \mathcal{P}(X_d) = \mathcal{F}(X_d), \tag{4}$$

where $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ indicate the amplitude and phase respectively. To reduce the effect of altered amplitude component,

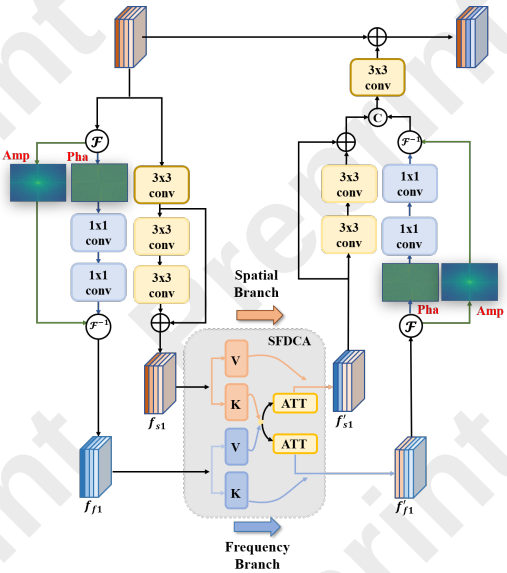

Figure 3: Phase-emphasized interaction block (PEI). Space-frequency dual cross attention (SFDCA) is shown in Figure 4.

we recombine to obtain $G'_r \in \mathbb{R}^{H_s\times W_s\times C_s}$:

$$G'_r = \mathrm{Conv}(\mathcal{F}^{-1}(\mathcal{A}(X_d), \mathcal{P}(G_r)) + G_r), \tag{5}$$

where Conv denotes a 1×1 convolution layer. Following [Chen *et al.*, 2021b; Wang *et al.*, 2022b], we also use CNNs to extract edges from coarse features as a kind of supervised information. Meanwhile, the coarse features G'_r are adaptively divided into low-high frequency for learning separately through a predicted frequency demarcation. Then, the features after the frequency separation subnetwork are fused with spatial information to output the predicted localization map. Finally, we pool the predicted probability map and use a fully connected layer for forgery detection.

3.3 Spectral decomposition subnetwork

Most forged images are carefully processed to hide tampering artifacts, making it challenging to model inconsistencies in the spatial domain. To overcome this challenge, SDSN employs deep Fourier transforms to independently process the

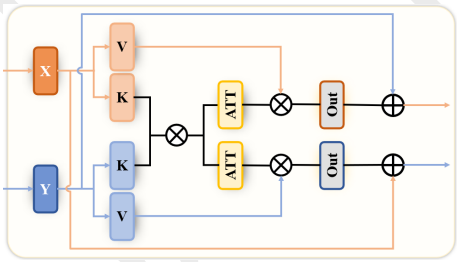

Figure 4: Space-frequency dual cross attention (SFDCA).

amplitude and phase spectra of images, with a heightened focus on phase learning.

We use ResNet-50 pretrained on ImageNet [Deng *et al.*, 2009] as the backbone network of the spectral decomposition subnetwork. To enhance the phase characterization, we design the phase-emphasized interaction (PEI) block and place it alternately with the ResNet block, as shown in Figure 3. The PEI blocks specifically learn the phase information of forged images, and progressively improve the network's ability to capture the tampered artifacts, thus improving the accuracy of the forgery localization.

We illustrate the PEI block as shown in Figure 3. According to spectral convolution theorem in Fourier theory [Katznelson, 2004], processing information in Fourier space is capable of capturing the global frequency representation in the frequency domain. In contrast, the normal convolution focuses on learning local representations in the spatial domain. Thus, we introduce the dual domain information interaction to facilitate the information flow and learn the complementary representation. Specifically, it comprises a spatial branch and a frequency branch for processing spatial and frequency representations. Denoting $G_p \in \mathbb{R}^{H_p \times W_p \times C_p}$ as the input features of the PLB block, the frequency branch first uses a 1×1 convolution to process G_p that obtains G_{f_0} and then adopts Fourier transform to convert it to the Fourier space:

$$\mathcal{A}\left(G_{f_0}\right), \mathcal{P}\left(G_{f_0}\right) = \mathcal{F}\left(G_{f_0}\right). \tag{6}$$

Next, we adopt the operation $\mathcal{OF}(\cdot)$ that consists of 1×1 convolution layers on its phase component, and then recompose the operated result with the amplitude component to obtain $G_{f_1} \in \mathbb{R}^{H_p \times W_p \times C_p}$, which is expressed as:

$$G_{f_1} = \mathcal{F}^{-1}\left(\mathcal{A}(G_{f_0}), \mathcal{OF}(\mathcal{P}(G_{f_0}))\right), \tag{7}$$

In this way, G_{f_1} is the processed result of the frequency-domain representation with enhanced phase information. Meanwhile, the spatial branch processes information in the spatial domain to obtain $G_{s_1} \in \mathbb{R}^{H_p \times W_p \times C_p}$:

$$G_{s_1} = \mathcal{OS}\left(\left(G_p\right)\right), \tag{8}$$

where $\mathcal{OS}$ denotes a residual block with 3×3 convolution layers. Then, inspired by [Luo *et al.*, 2021], we introduce the space-frequency dual cross attention (SFDCA) to interact with frequency domain features and spatial features. As Figure 4 illustrates, SFDCA employs shared attention to facilitate the comprehensive interplay between the two domains, a process that can be described as:

$$G'_{f_1}, G'_{s_1} = f_{\mathrm{SFDCA}}(G_{f_1} + G_{s_1}), \tag{9}$$

where G'_{f_1} and G'_{s_1} are the output of the interacted spatial branch and frequency branch. And they both get the complementary representation, which benefits for these two branches to obtain more representational features. The following spatial and frequency branches are formulated in the same way as above and output G_{f_2} and G_{s_2}:

$$G_{f_2} = \mathcal{F}^{-1}\left(\mathcal{A}(G'_{f_1}), \mathcal{OF}(\mathcal{P}(G'_{f_1}))\right), \tag{10}$$

$$G_{s_2} = \mathcal{OS}\left(\left(G'_{s_1}\right)\right). \tag{11}$$

Finally, we concatenate G_{f_2} and G_{s_2} and then apply a 1×1 convolution operation to integrate them:

$$G_c = \mathrm{Conv}\left(\mathrm{Cat}(G_{f_2}, G_{s_2})\right), \tag{12}$$

where $G_c \in \mathbb{R}^{H_p \times W_p \times C_p}$ is the output of PEI block. The phase information of the forged images is effectively enhanced by the subnetwork with five PEI blocks allowing to obtain features containing more forgery traces.

3.4 Frequency separation subnetwork

As described in Sec. 3.2, the coarse features G'_r are fed into adaptive frequency separation subnetwork. It adaptively divides the forged image into low and high frequencies components with different gradients respectively, and uses the spatial branch for information compensation.

$G'_r \in \mathbb{R}^{H_s \times W_s \times C_s}$ is first transformed into frequency domain by Fourier Transform:

$$\mathcal{A}\left(G'_r\right), \mathcal{P}\left(G'_r\right) = \mathcal{F}\left(G'_r\right). \tag{13}$$

Then, the subnetwork predicts a two-dimensional mask with mask prediction network (MPN). Specifically, the G'_r is mapped into one-dimensional vector by global average pooling and then pass through the fully-connected layers to generate two scalars α and β in the range of 0 to 1:

$$\alpha, \beta = \sigma(\mathrm{FC}(\mathrm{GAP}(G'_r))), \tag{14}$$

where σ denotes the sigmoid activation function. The mask $M \in \mathbb{R}^{H_s \times W_s}$ is obtained by setting the corresponding bounding box as 1 and the remaining as 0:

$$M\left[\frac{H}{2} - \frac{\alpha}{2}H : \frac{H}{2} + \frac{\alpha}{2}H, \frac{W}{2} - \frac{\beta}{2}W : \frac{W}{2} + \frac{\beta}{2}W\right] = 1. \tag{15}$$

Then, based on the obtained mask, we filter out the low and high frequencies parts of the coarse features G'_r. Meanwhile, the convolution operation is performed over amplitude to enhance the learning ability of the subnetwork:

$$\begin{aligned} G_L &= \mathcal{F}^{-1}\left(M \odot \mathcal{A}\left(G'_r\right), \mathcal{OF}\left(\mathcal{P}\left(G'_r\right)\right)\right), \\ G_H &= \mathcal{F}^{-1}\left((1-M) \odot \mathcal{A}\left(G'_r\right), \mathcal{OF}\left(\mathcal{P}\left(G'_r\right)\right)\right), \end{aligned} \tag{16}$$

where $\odot$ is the Hadamard product, $\mathcal{OF}(\cdot)$ denotes two 1×1 convolution layers, G_L and G_H denote the low frequency and high frequency components, respectively. Next, the two components are processed with convolution layers separately in the spatial domain and fused:

$$G_F = \mathrm{Conv}\left(\mathrm{Cat}\left(\mathcal{OS}\left(G_L\right), \mathcal{OS}\left(G_H\right)\right)\right), \tag{17}$$

where $G_F \in \mathbb{R}^{H_s \times W_s \times C_s}$ is the overall frequency features, $\mathcal{OS}$ denotes a residual block with 3×3 convolution layers

Loc.	Data	Col.	Cov.	CAS.	NI.16	IM.20
		Metric: AUC(%) – Pre-trained				
ManTra	64K	82.4	81.9	81.7	79.5	74.8
SPAN	96k	93.6	92.2	79.7	84.0	75.0
PSCC	100k	98.2	84.7	82.9	85.5	80.6
Ob.Fo.	62K	95.5	92.8	84.3	87.2	82.1
TANet	60K	**98.7**	91.4	85.3	<u>89.8</u>	<u>84.9</u>
HiFi	100k	98.3	<u>93.2</u>	<u>85.8</u>	87.0	82.9
Ours	60K	<u>98.5</u>	**94.2**	**91.0**	**89.9**	**85.2**

(a)

Loc.	Cov.	CAS.	NI.16
	Metric: AUC(%) / F1(%) – Fine-tuned		
RGB-N	81.7/43.7	79.5/40.8	93.7/72.2
SPAN	93.7/55.8	83.8/38.2	96.1/58.2
PSCC	94.1/72.3	87.5/55.4	99.6/81.9
Ob.Fo.	95.7/75.8	88.2/57.9	99.6/82.4
TANet	<u>97.8</u>/78.2	<u>89.3</u>/ 61.4	<u>99.7</u>/<u>86.5</u>
HiFi	96.1/<u>80.1</u>	88.5/<u>61.6</u>	98.9/85.0
Ours	**98.3/81.5**	**92.6/65.3**	**99.8/87.2**

(b)

Det.	AUC(%)	F1(%)
ManTra	59.94	56.69
SPAN	67.33	63.48
PSCC	99.65	97.12
Ob.Fo.	<u>99.70</u>	97.34
HiFi	99.50	<u>97.40</u>
Ours	**99.73**	**97.52**

(c)

Table 1: Image forgery detection and localization results. (a) Localization performance of the pre-train model. (b) Localization performance of the fine-tuned model. (c) Detection performance on *CASIA-D* dataset. (Bold means best, underline means second best).

and Conv is the 1×1 convolution. Besides, we also adopt a residual block to process the spatial branch. For the fusion of the two branches, We follow [Chen *et al.*, 2021b] to adopt the Dual Attention (DA) module [Fu *et al.*, 2019]. DA includes both channel attention and position attention. It can effectively fuse two branches. The process can be written as:

$$G_S = \mathcal{OS}\left(G_r'\right), \tag{18}$$

$$G_o = \mathrm{DA}\left(G_F, G_S\right). \tag{19}$$

Then, we transform $G_o \in \mathbb{R}^{H_s \times W_s \times 1}$ with bilinear upsampling into the final predicted mask $G_{out} \in \mathbb{R}^{H \times W \times 1}$. For the detector, we apply the ConvGeM proposed by MVSS-Net++ [Dong *et al.*, 2023], which can convert localization results G_{out} into detection prediction D_{out}. ConvGeM strikes a good balance between detection and localization through a decayed skip connection. Thus, we use ConvGem to obtain a more accurate detection result:

$$D_{out} = ConvGeM(G_{out}) \tag{20}$$

3.5 Optimization

Following most studies [Chen *et al.*, 2021b; Salloum *et al.*, 2018; Wang *et al.*, 2022b], we also employ the edge supervision. However, this is not the focus of this work, so we have used some common methods. Following [Chen *et al.*, 2021b], we use the Sobel layer and the residual block to obtain the edge prediction $G_e \in \mathbb{R}^{H_e \times W_e \times 1}$ in a shallow-to-deep manner. For edge loss, the ground-truth edges $E \in \mathbb{R}^{H \times W \times 1}$ is downsampled to a smaller size $E' \in \mathbb{R}^{H_e \times W_e \times 1}$ to match G_e. This strategy outperforms upsampling G_e in terms of computational cost and performance. The overall loss function can be written as:

$$\mathbf{L} = \alpha\mathcal{L}_1(Y, G_{out}) + \beta\mathcal{L}_2(y, D_{out}) + (1-\alpha-\beta)\mathcal{L}_3(E', G_e), \tag{21}$$

where $\mathcal{L}_1$ and $\mathcal{L}_3$ denote the Dice loss [Chen *et al.*, 2021b], $\mathcal{L}_2$ is BCE loss, y is a label that represents the authenticity of the image, $Y \in \mathbb{R}^{H \times W \times 1}$ is the ground-truth mask, and α, β are the hyperparameters to balance the loss function. In practice, α is set as 0.60 and β is set as 0.2. Note that authentic images are only used to compute $\mathcal{L}_2$.

4 Experiments

4.1 Experimental Setup

Pre-training Data We create a sizable image tampering dataset and use it to pre-train our model. This dataset includes three categories: 1) splicing, 2) copy-move, and 3) removal. We ensure that the training and test datasets are dissimilar.

Testing Datasets Following [Liu *et al.*, 2022; Wang *et al.*, 2022a], we evaluate our model on CASIA [Dong *et al.*, 2013], Coverage [Wen *et al.*, 2016], Columbia [Hsu and Chang, 2006], NIST16 [Guan *et al.*, 2019] and IMD20 [Novozamsky *et al.*, 2020]. Specifically, IMD20 collects real-life manipulated images from Internet. We apply the same training/testing splits as [Hu *et al.*, 2020; Wang *et al.*, 2022a] to fine-tune our model for fair comparisons.

4.2 Image Forgery Localization

Following SPAN [Hu *et al.*, 2020], our model is compared with other state-of-the-art tampering localization methods under two settings: 1) training on the synthetic dataset and evaluating on the full test datasets, and 2) fine-tuning the pre-trained model on the training split of test datasets and evaluating on their test split. The pre-trained model demonstrates generalizability, while the fine-tuned model shows local performance after reducing domain discrepancy.

Pre-trained Model Table 1a shows the localization performance of pre-trained models for different methods on five datasets under pixel-level AUC. We compare our model F2D-Net with MantraNet [Wu *et al.*, 2019], SPAN [Hu *et al.*, 2020], PSCCNet [Liu *et al.*, 2022], ObjectFormer [Wang *et al.*, 2022a], TANet [Shi *et al.*, 2023], and HiFi_IFDL [Guo *et al.*, 2023] when evaluating pre-trained models. The pre-trained F2D-Net achieves the best localization performance on Coverage, CASIA, and NIST16, IMD20 and ranks the second on Columbia. Especially, F2D-Net achieves 94.2 % on the copy-move dataset COVER, whose image forgery regions are indistinguishable from the background. This validates our model owns the superior ability to capture tampering traces. We fail to achieve the best performance on Columbia, despite surpassing TANet 0.2 % under AUC. We contend that the explanation may be that the distribution of their synthesized training data closely resembles that of the Columbia dataset. This is further supported by the results in Table 1b, which show that F2D-Net performs better than TANet in terms of both AUC and F1 scores. Furthermore, it is worth pointing out F2D-Net achieves decent results with less pre-training data.

Fine-tuned Model The network weights of the pretrained model are used to initiate the fine-tuned models that will be trained on the training split of Coverage, CASIA, and NIST16 datasets, respectively. We evaluate the fine-tuned models of

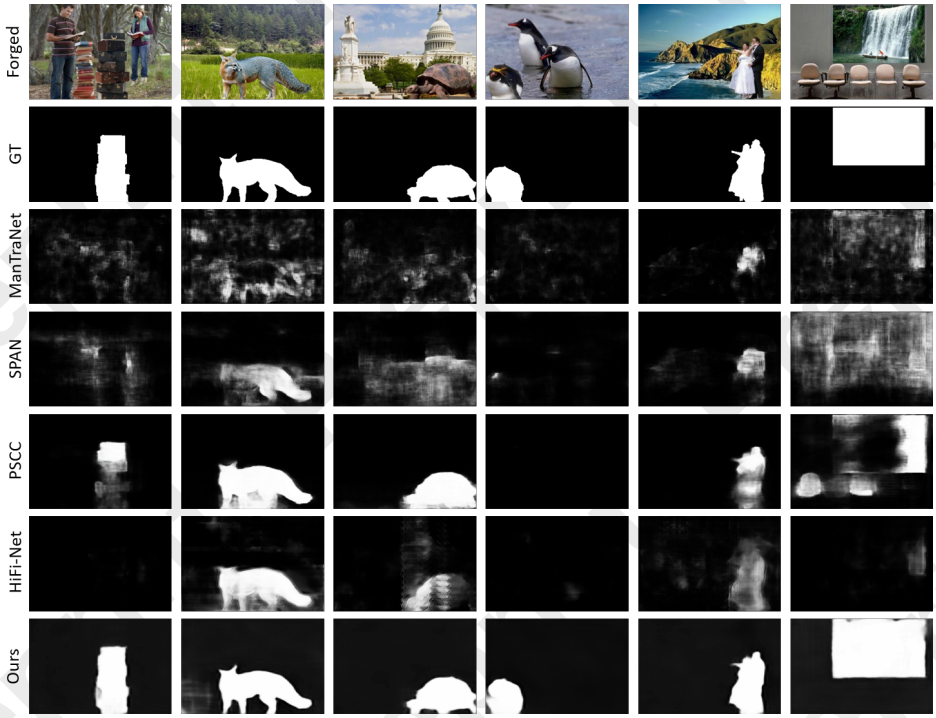

Figure 5: Visualization of the predicted manipulation mask by different methods. From top to bottom, we show forged images, GT masks, predictions of ManTraNet, SPAN, PSCC-Net, HiFi-Net and ours.

Distortion	SPAN	ObjectFormer	Ours	
no distortion	83.95	87.18	**89.89**	
Resize(0.78×)	83.24	87.17	**89.76**	**0.13**↓
Resize(0.25×)	80.32	86.33	**88.17**	**1.72**↓
Blur($k = 3$)	83.10	85.97	**89.55**	**0.34**↓
Blur($k = 15$)	79.15	80.26	**87.94**	**1.95**↓
Noise($\sigma = 3$)	75.17	79.58	**88.63**	**1.26**↓
Noise($\sigma = 15$)	67.28	78.15	**83.58**	**6.31**↓
Compress($q = 100$)	83.59	86.37	**89.80**	**0.09**↓
Compress($q = 50$)	80.68	86.24	**89.21**	**0.68**↓

Table 2: Localization performance on NIST16 dataset under various distortions. AUC scores are reported (in %), (Blur: GaussianBlur, Noise: GaussianNoise, Compress: JPEGCompress.)

different methods in Table 1b. As for AUC and F1, our model achieves significant performance gains. This validates that F2D-Net could precisely capture subtle tampering traces by phase leaning and adaptive low-high frequency learning.

4.3 Image Forgery Detection

To avoid false alarms, we engage in forgery detection tasks. Following the ObjectFormer [Wang *et al.*, 2022a], we conduct experimental comparisons on the CASIA-D dataset introduced by [Liu *et al.*, 2022]. As shown in Table 1c, our method delivers exceptional detection performance, with an AUC of 99.73% and an F1 score of 97.52%. Our method learns to differentiate and conquer forgery features in the frequency domain, which allows for the distinct separation of forged images from authentic ones.

Variants	CASIA		NIST16	
	AUC	F1	AUC	F1
baseline	70.5	36.7	76.2	51.6
w/o ALP	77.3	49.2	87.4	62.9
w/o SFDCA	87.1	52.8	95.6	78.3
w/o FSSN	88.5	53.9	97.2	80.1
Ours	**92.6**	**65.3**	**99.8**	**87.2**

Table 3: Ablation results on CASIA and NIST16 dataset using different variants of F2D-Net. AUC and F1 scores (%) are reported.

4.4 Robustness Evaluation

To analyze the robustness of F2D-Net for localization, we follow the distortion settings in [Wang *et al.*, 2022a] to degrade the raw forged images from NIST16. These distortions types include resizing images to different scales (Resize), applying Gaussian blur with a kernel size k (GaussianBlur), adding Gaussian noise with a standard deviation σ (GaussianNoise), and performing JPEG compression with a quality factor q (JPEGCompress). We compare the forgery localization performance (AUC scores) of our pretrained models with SPAN[Hu *et al.*, 2020] and ObjectFormer on these corrupted data, and report the results in Table 2. F2D-Net demonstrates better robustness against various distortion techniques. It is worth noting that JPEG compression is commonly performed when uploading images to social media. Our model outperforms others on compressed images.

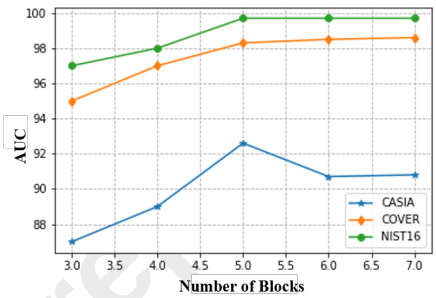

Figure 6: AUC score of our framework with different numbers of the PEI blocks.

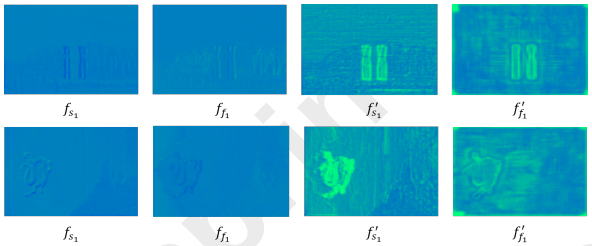

Figure 7: Feature visualization of different representations in the PEI block.

4.5 Ablation Analysis

In this section, we conduct experiments to demonstrate the effectiveness of our method F2D-Net. More ablation studies are provided in supplementary materials. The F2D-Net contains two key components: spectral decomposition subnetwork with the phase-emphasized interaction (PEI) block and frequency separation subnetwork. The PEI block comprises two designs: one is an enhanced learning dedicated to phase spectra, termed as the additional learning of phase (ALP), and the other is the space-frequency dual cross attention (SFDCA) used for spatial-frequency interaction. The frequency separation subnetwork (FSSN) is designed to decompose the tampering traces into low and high frequencies. To evaluate the effectiveness of ALP, SFDCA and FSSN, we remove them separately from F2D-Net and evaluate the forgery localization performance on CASIA and NIST16 datasets.

Table 3 presents the quantitative outcomes. The baseline denotes that we just use ResNet-50 and ResBlock. It can be observed that without FSSN, the AUC scores decrease by 4.4 % on CASIA and 2.5 % on NIST16, while without SFDCA, the AUC scores decrease by 5.9 % on CASIA and 4.1 % on NIST16. Moreover, when ALP is discarded, significant performance degradation in Table 3, i.e., 16.5 % in terms of AUC and 24.6 % in terms of F1 on CASIA can be observed.

In Figure 6, we show the different numbers of the phase-emphasized interaction (PEI) block to verify its effect over three datasets. There is an overall incremental trend in the forger location performance as the number of PEI increases, but the performance saturates after reaching a critical point. It is obvious that the setting of 5 is the optimal solution.

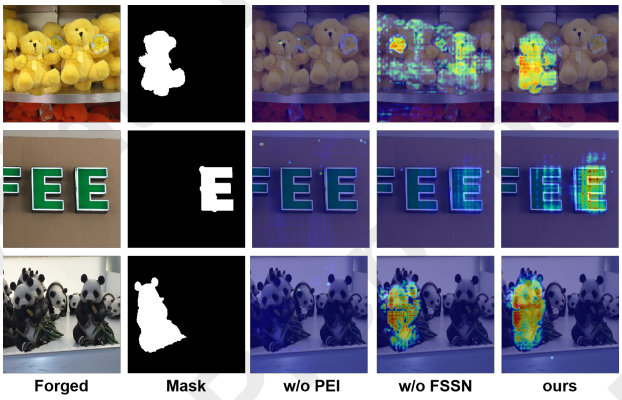

Figure 8: Visualization of our framework. From left to right, we display the forged images, masks, GradCAM of the feature map without (w/o) and with (w) PEI and FSSN.

4.6 Visualization Results

Qualitative results. We provide predicted forgery masks of different methods in Figure 5. Since the source code of ObjectFormer [Wang *et al.*, 2022a] is not available, their predictions are not available. It benefits from the ability of our model to capture subtle tampering traces and decompose them into low and high frequencies.

Visualization of PEI. To verify the effect of the phase-emphasized interaction (PEI) block, we show the feature visualization of different representations in the PEI block in Figure 7. As can be seen, since features after interaction can obtain complementary representations from each other, f'_{f_1} obtains more spatial information and the details in f'_{s_1} are enhanced. It benefits for these two branches to obtain more representational features. Besides, we show the change of features before and after PEI in Figure 8. It is clear that PEI enforces the forgery features learning.

Visualization of FSSN. To verify the usefulness of the frequency separation subnetwork (FSSN), the change of features before and after FSSN is shown in Figure 8. The results demonstrate that the FSSN can effectively refine forgery localization and prevents false alarms, thus helping our model to perform well.

5 Conclusion

In this paper, we explore image forgery from a frequency perspective and propose a novel cascade frequency learning network for Image forgery detection and localization. In detail, we first adopt deep Fourier transform and introduce a phase-emphasized interaction block to learn the phase information for capturing manipulation traces precisely. Then, we design an adaptive frequency separation subnetwork to decompose the tampering traces into low and high frequencies, realizing forgery localization refinement. Extensive experimental results on several benchmarks demonstrate the effectiveness of the proposed algorithm.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008, 62422609, 62276243, and 62476260, and by the Fundamental Research Funds for the Central Universities under Grant WK2100000057.

References

[Aloraini *et al.*, 2020] Mohammed Aloraini, Mehdi Sharifzadeh, and Dan Schonfeld. Sequential and patch analyses for object removal video forgery detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):917–930, 2020.

[Bammey *et al.*, 2020] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204, 2020.

[Chen *et al.*, 2021a] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1081–1088, 2021.

[Chen *et al.*, 2021b] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14185–14193, 2021.

[Cozzolino *et al.*, 2015] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, 10(11):2284–2297, 2015.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Dong *et al.*, 2013] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426. IEEE, 2013.

[Dong *et al.*, 2023] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2023.

[Frigo and Johnson, 1998] Matteo Frigo and Steven G Johnson. Fftw: An adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, pages 1381–1384. IEEE, 1998.

[Fu *et al.*, 2019] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Guan *et al.*, 2019] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72. IEEE, 2019.

[Guo *et al.*, 2023] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3155–3165, June 2023.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hsu and Chang, 2006] J Hsu and SF Chang. Columbia uncompressed image splicing detection evaluation dataset. *Columbia DVMM Research Lab*, 2006.

[Hu *et al.*, 2020] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *European conference on computer vision*, pages 312–328. Springer, 2020.

[Huh *et al.*, 2018] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018.

[Katznelson, 2004] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

[Liu *et al.*, 2022] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[Luo *et al.*, 2021] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021.

[Novozamsky *et al.*, 2020] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In

Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, pages 71–80, 2020.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. *URL https://arxiv. org/abs/2204.06125*, 7, 2022.

[Salloum *et al.*, 2018] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, 51:201–209, 2018.

[Shi *et al.*, 2023] Zenan Shi, Haipeng Chen, and Dong Zhang. Transformer-auxiliary neural networks for image manipulation localization by operator inductions. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.

[Wang *et al.*, 2022a] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022.

[Wang *et al.*, 2022b] Menglu Wang, Xueyang Fu, Jiawei Liu, and Zheng-Jun Zha. Jpeg compression-aware image forgery localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5871–5879, 2022.

[Wen *et al.*, 2016] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*, pages 161–165. IEEE, 2016.

[Wu *et al.*, 2019] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019.

[Xu *et al.*, 2021] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021.

[Zhou *et al.*, 2018] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061, 2018.