

FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing

Hossein Kashiani Niloufar Alipour Talemi Fatemeh Afghah
 Clemson University
 {hkashia, nalipou, fafghah}@clemson.edu

Abstract

Deepfake detectors often struggle to generalize to novel forgery types due to biases learned from limited training data. In this paper, we identify a new type of model bias in the frequency domain, termed *spectral bias*, where detectors overly rely on specific frequency bands, restricting their ability to generalize across unseen forgeries. To address this, we propose *FreqDebias*, a frequency debiasing framework that mitigates spectral bias through two complementary strategies. First, we introduce a novel *Forgery Mixup (Fo-Mixup)* augmentation, which dynamically diversifies frequency characteristics of training samples. Second, we incorporate a dual consistency regularization (CR), which enforces both local consistency using class activation maps (CAMs) and global consistency through a von Mises-Fisher (vMF) distribution on a hyperspherical embedding space. This dual CR mitigates over-reliance on certain frequency components by promoting consistent representation learning under both local and global supervision. Extensive experiments show that *FreqDebias* significantly enhances cross-domain generalization and outperforms state-of-the-art methods in both cross-domain and in-domain settings.

1. Introduction

Recent advances in face forgery enable malicious users to create ultrarealistic facial images that appear genuine, which poses threats to societal trust and security [1, 14, 27, 33]. While detection methods have shown progress, they often struggle to generalize beyond specific in-domain forgery types, limiting their effectiveness in detecting out-of-distribution samples [6, 7, 12, 15, 24, 30, 34, 38, 49, 71]. This limitation arises from model bias, where detectors overly rely on spurious correlations, such as identity [24], background [34], specific-forgery [63, 65], or structural information [9], rather than common forgery features shared across various forgeries. While recent studies [9, 13, 24, 63, 65] have focused on addressing human-perceptible bias, to

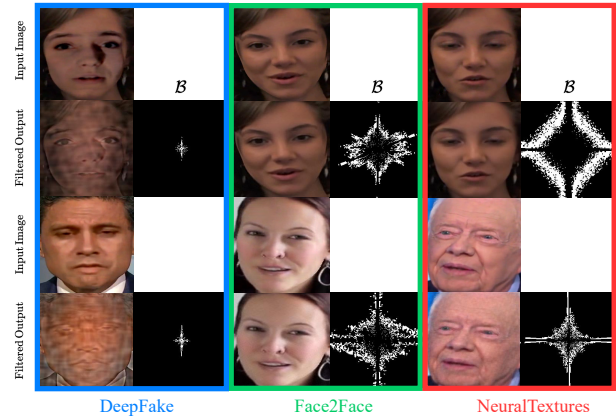


Figure 1. Visualization of certain frequency bands overly relied upon by the vanilla deepfake detector. The dominant frequency bands, highlighted in binary masks $\mathcal{B}$, are restricted to specific bands and also driven by the forgery type. Further details are provided in the supplementary material.

the best of our knowledge, imperceptible model bias in the frequency domain remains unexplored. An analysis in this domain can provide insights into the causes of this bias and support mitigation efforts to enhance generalization.

In this paper, we investigate model bias in the frequency domain and evaluate its impact on the generalization capabilities of deepfake detectors. To this end, we design a spectral bias experiment to identify the range of frequency bands that a standard deepfake detector pays attention to during classification. Our analysis reveals that detectors often over-rely on certain frequency bands, termed dominant frequency components, while failing to capture general forgery artifacts that may be spread across wider frequency bands. These components are defined as bands whose exclusion leads to the largest increase in classification loss, indicating the detector’s critical dependence. As shown in Fig. 1, these dominant frequency components are primarily driven by the forgery type, each exhibiting unique statistical patterns and typically confined to specific frequency bands. This over-reliance limits the generalization capac-

ity of detectors across unseen forgeries. Our findings align with prior study [43] showing that neural networks prioritize dominant frequency components in the statistical structure of training data to reduce optimization loss, at the cost of capturing semantic features necessary for generalization.

Building on these observations, we introduce a frequency debiasing framework, termed FreqDebias, to mitigate spectral bias and improve generalization of our detector beyond in-domain forgeries. FreqDebias incorporates two complementary strategies to achieve this objective. First, we introduce a novel Forgery Mixup (Fo-Mixup) which diversifies the frequency characteristics of training samples to address the over-reliance of our detector on dominant frequency components. Fo-Mixup dynamically modulates the amplitude spectra within these dominant frequency bands, creating diverse and challenging forgery samples to broaden the exposure of our detector to varied frequency bands.

Second, to explicitly regularize the learning behavior of our detector across various frequency components, we propose a dual consistency regularization (CR) that enforces our detector to learn consistent representation in both local and global supervision. For local supervision, the sensitivity of the last convolution layer w.r.t. the spatial distribution of discriminative patterns is regularized using class activation maps (CAMs) [72]. Since Fo-Mixup preserves the spatial location of tampered regions, enforcing CAM consistency ensures that our forgery detector remains focused on the discriminative regions, making it less susceptible to domain-specific noise or other forgery-irrelevant artifacts that may arise from dominant frequency components.

Moreover, for global supervision, we model facial feature representations using the von Mises-Fisher (vMF) distribution to impose CR on a hyperspherical embedding space. Compared with the Gaussian distribution, the vMF distribution exploits the geometric properties of hyperspherical space to learn a global view of the embedding space. [22, 29, 32, 57]. Also, the vMF distribution can better capture semantic information where embedding directions specify semantic classes [48]. Leveraging this property, we impose CR on this unit hypersphere to facilitate global, geometrically consistent learning as shown in Fig. 2. To substantiate the generalization of our method, we conduct extensive experiments under diverse face forgery attacks. The results validate that our FreqDebias framework significantly outperforms the state-of-the-art (SOTA) studies under the cross-domain setting while maintaining its competitive performance under the in-domain setting.

Our main contributions are summarized as follows:

1. We discover spectral bias in deepfake detection and propose the Fo-Mixup augmentation to mitigate this bias. Fo-Mixup exposes the detector to a wider range of frequencies to reduce over-reliance on specific frequency bands, thus enhancing its generalization capabilities.
2. We introduce a dual consistency regularization (CR) to explicitly regulate our detector by jointly enforcing local and global consistency. By leveraging CAMs for local consistency to maintain focus on discriminative regions and the vMF distribution for global consistency on a hyperspherical embedding space, our dual CR effectively mitigates the spectral bias.
3. Extensive experiments in both in-domain and cross-domain settings demonstrate the effectiveness and generalization of our framework compared to existing studies.

2. Related Works

Deepfake Detection. In response to the growing threat of face forgery attacks, numerous studies have advanced forgery detection techniques [18, 38, 49]. Some studies focus on capturing forgery artifacts in the spatial [31, 49] or frequency [36, 38] domains. For instance, Qian et al. [46] utilize local frequency statistics via the Discrete Cosine Transform, while Liu *et al.* [36] apply the Discrete Fourier Transform to detect up-sampling artifacts. However, their performance degrades significantly on unseen forgeries. To address generalization, recent studies [19, 24, 38, 63] propose various strategies. One method, CORE [41], addresses overfitting by enforcing representation consistency across pairs of standard augmentations to enhance generalization. However, CORE limits the potential of CR by restricting augmentation diversity. In contrast, our method synthesizes challenging forgeries to address the spectral bias in deepfake detection to learn more generalizable representations.

Debiasing and Spurious Correlations. Recent studies have demonstrated that deep learning models are vulnerable to spurious correlations [16, 35]. These are irrelevant features linked to specific classes that harm generalization [17]. In deepfake detection, such biases emerge as dependence on unintended artifacts like identity [24], background [34], structural information [9], or forgery-specific patterns [63, 65], degrading performance on unseen forgeries. Additionally, such correlations can emerge in the frequency domain, driving models to exploit spectral shortcuts [26, 58]. Data augmentation is an effective strategy to counteract these biases [45, 67]. While prior studies [25, 61, 66, 68, 70] have explored Fourier-based methods, they overlook the critical role of dominant frequencies. Our method explicitly disrupts dominant frequencies to mitigate spurious shortcuts and improve generalization.

Von Mises-Fisher Distribution. The vMF distribution is a probability distribution function for directional data which can statistically model samples with unit norm. An important advantage of the vMF distribution over the Gaussian distribution is its ability to model directional data with a hyperspherical interpretation. This ability makes it more appropriate as a prior for directional data. Several recent

studies in computer vision [22, 32, 48, 57, 60] have highlighted the effectiveness of the vMF distribution in hyperspherical representation learning. In [29], the vMF distribution regularizes intra-class feature distribution to improve performance on imbalanced data. Kirchhof *et al.* [28] introduce probabilistic proxy-based deep metric learning using directional vMF distributions to handle image-intrinsic uncertainties. Wang *et al.* [57] extend cosine-based classifiers with a vMF mixture model, enabling a quantitative quality measurement in hypersphere space for long-tailed learning. In this study, we utilize the vMF distribution to globally enhance CR across diverse face forgery variations, ensuring robust generalization.

3. Methodology

Given an original training dataset $T = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$ with N_t samples, where each x_i^t and y_i^t represent the i -th image and its corresponding class label, we employ Fo-Mixup to generate a diverse set of synthetic forgery images. Additionally, we use standard data augmentation to produce augmented versions of the real images. This process yields a synthesized dataset $S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ comprising N_s samples, where x_i^s and y_i^s denote the i -th synthetic image and its class label, respectively. Using both the synthesized forgery images and the augmented real images, we then regularize our model to maintain consistency across the original T and synthesized S domains for both real and forgery classes.

3.1. Forgery Mixup Augmentation

The amplitude spectrum of an image retains low-level statistics, such as forgery artifacts, color, style information, and sensor artifacts, while the phase spectrum includes high-level semantic content [42, 44, 61]. In deepfake detection, training the detector without specific considerations can cause the model to over-rely on domain-specific details (such as hair color, background, texture, and style) encoded in the amplitude spectrum [38]. Our experiments show that this overemphasis on dominant frequency components can lead to overfitting, where the model performs well on training data but poorly on out-of-distribution forgeries.

To address these challenges, we introduce the Fo-Mixup augmentation to modulate the dominant frequency components within the amplitude spectrum of different forgery images. Fo-Mixup exposes our forgery detector to more variations in the dominant frequency components of the amplitude spectrum during training. This exposure encourages the detector to attend to a broader range of frequency bands, thereby promoting the learning of more generalizable features and reducing overfitting.

Given an input forgery image x_i , we apply the fast Fourier transformation $\mathcal{F}(x_i)$ on each RGB channel to compute its amplitude $\mathcal{A}(x_i)$ and phase $\mathcal{P}(x_i)$ components. Fo-Mixup initiates with the identification of the key frequency

Algorithm 1 Fo-Mixup Augmentation

Inputs: Input forgery images x_i and x_j , interpolation hyperparameter ξ , total number of angular segments T , total number of clusters k , top number of clusters t , constant distance r , constant angle θ from the horizontal axis, a normal distribution $p_{\mathcal{A}}$, pre-trained forgery detector.

Outputs: Augmented forgery image x_{ij}

- 1: $[\mathcal{A}(x_i), \mathcal{P}(x_i)] \leftarrow \mathcal{F}(x_i)$ ▷ FFT for x_i
- 2: $[\mathcal{A}(x_j), \mathcal{P}(x_j)] \leftarrow \mathcal{F}(x_j)$ ▷ FFT for x_j
- 3: $\{\mathcal{A}_{\text{seg}_1}, \dots, \mathcal{A}_{\text{seg}_T}\} \leftarrow \text{Partition}(\mathcal{A}(x_i, r, \theta))$ ▷ Partition $\mathcal{A}(x_i)$ into T angular segments with constant distance r and angle θ
- 4: $\{\mu_{\text{seg}_1}, \dots, \mu_{\text{seg}_T}\} \leftarrow \{\frac{1}{N_{\text{seg}_z}} \sum_{\mathcal{A}_{\text{seg}_z}} \log(1 + |\mathcal{A}_{\text{seg}_z}|)\}_{z=1}^T$ ▷ Mean log-scaled spectrum
- 5: $\{\mu_1, \dots, \mu_k\} \leftarrow \text{K-means}(\{\mu_{\text{seg}_1}, \dots, \mu_{\text{seg}_T}\})$ ▷ Clustering
- 6: $\{\mathcal{B}_1, \dots, \mathcal{B}_k\} \leftarrow \{\text{indices}_{\mu_z}\}_{z=1}^k$ ▷ Binary masks
- 7: $\{\mathcal{A}_{\mu_1}, \dots, \mathcal{A}_{\mu_k}\} \leftarrow \{\mathcal{A}(x_i) \otimes \mathcal{B}_z\}_{z=1}^k$ ▷ Filtering
- 8: $\{x_{\mu_1}, \dots, x_{\mu_k}\} \leftarrow \{\mathcal{F}^{-1}[\mathcal{A}_{\mu_z}, \mathcal{P}(x_i)]\}_{z=1}^k$ ▷ Filtered images
- 9: $\{x_{\mu_{s1}}, \dots, x_{\mu_{st}}\} \leftarrow \text{OHM}(\{x_{\mu_1}, \dots, x_{\mu_k}\})$ ▷ Apply OHM [50] to the filtered images using our pre-trained forgery detector
- 10: $\{\mathcal{B}_1, \dots, \mathcal{B}_t\} \leftarrow \{\text{indices}_{\mu_{sz}}\}_{z=1}^t$ ▷ Keep top t clusters, selected by OHM in step 9. These clusters correspond to the highest loss values.
- 11: $\mathcal{B} \leftarrow \text{RandomSelect}(\mathcal{B}_1, \dots, \mathcal{B}_t)$ ▷ Random selection
- 12: $\hat{\mathcal{A}}(x_{ij}) \leftarrow \mathcal{A}(x_i) \otimes \mathcal{B} + [(1 - \xi)\mathcal{A}(x_i) + \xi\mathcal{A}(x_j)] \otimes (1 - \mathcal{B})$
- 13: $x_{ij} \leftarrow \mathcal{F}^{-1}[p_{\mathcal{A}} \otimes \hat{\mathcal{A}}(x_{ij}), \mathcal{P}(x_i)]$ ▷ Reconstruct image
- 14: **Return** x_{ij} ▷ Augmented forgery image

components within the input forgery image.

To achieve this, we partition the amplitude spectrum into angular segments, with radial and angular components divided into intervals $r \in [i_r \Delta r, (i_r + 1) \Delta r]$ and $\theta \in [i_\theta \Delta \theta, (i_\theta + 1) \Delta \theta]$. The radial component measures distance from the origin, while the angular component defines orientation in the frequency domain. The indices i_r and i_θ denote discrete steps, and the step sizes Δr and $\Delta \theta$ determine the segmentation resolution. For each segment, we compute the mean logarithmic-scaled spectrum denoted as μ_{seg} , as the average of $\log(1 + |\mathcal{A}_{\text{seg}}|)$ values within the segment, where $\mathcal{A}_{\text{seg}}$ is the segment's amplitude spectrum.

Next, we apply the K-means clustering algorithm [39] to all angular segments using the mean spectrum μ_{seg} of all angular segments. The k cluster indices are then employed to construct k binary masks, which are subsequently used to filter the input image into k distinct filtered images. We feed the filtered images into our detector and arrange the classification losses in descending order, following the Online Hard Example Mining (OHM) [50]. Finally, we randomly select one of the top t binary masks (denoted as $\mathcal{B}$) as the ones associated with the highest loss values. To integrate the key frequency components of different forgery images, the selected binary mask $\mathcal{B}$ is applied to the amplitude component (shown in Fig. 2). These operations are formulated as follows:

$$\begin{aligned} \hat{\mathcal{A}}(x_{ij}) &= \mathcal{A}(x_i) \otimes \mathcal{B} \\ &+ [(1 - \xi)\mathcal{A}(x_i) + \xi\mathcal{A}(x_j)] \otimes (1 - \mathcal{B}), \quad (1) \end{aligned}$$

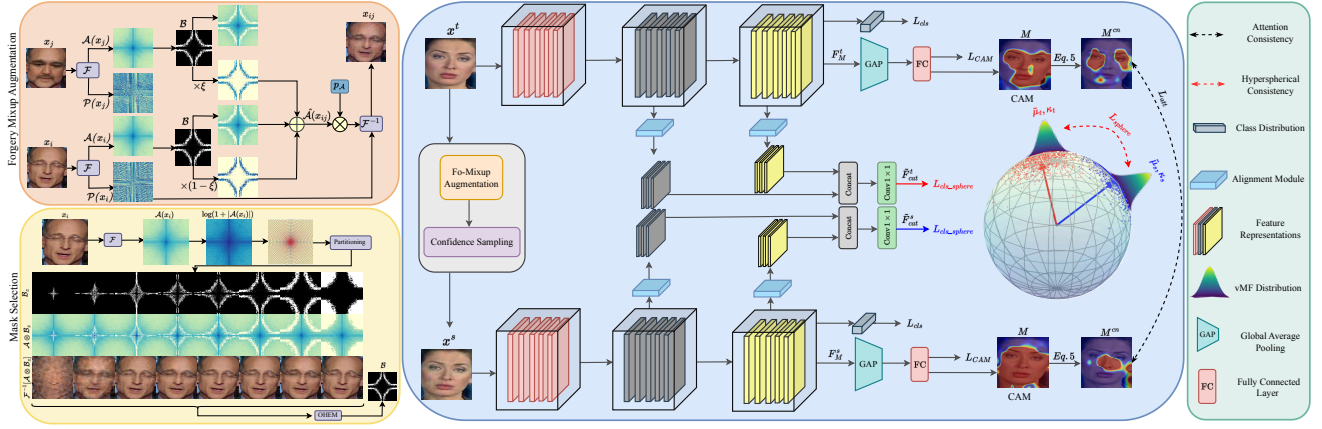


Figure 2. Overview of the proposed FreqDebias framework. First, Fo-Mixup diversifies the forgery class across frequency spectra by transforming x^t into x^s , while preserving the spatial structure of tampered regions. Then, the network is jointly regularized at local and global levels to enforce consistent representation learning. For the real class, the same process is applied using standard augmentation.

where x_i and x_j are the randomly selected forgery inputs, and ξ denotes the hyperparameter that adjusts the proportion of inputs. Also, $\otimes$ denotes element-wise multiplication. To expose our model in CR to more forgery artifacts variations in the dominant frequency component, we also perturb the amplitude spectrum as $p_A \otimes \hat{A}(x_{ij})$, where p_A is a normal distribution $\mathcal{N}(1, 0)$. Finally, the inverse Fourier transform reconstructs the newly stylized image using the phase component of the original input $\mathcal{P}(x_i)$ and the integrated amplitude spectrum $\hat{A}(x_{ij})$ as follows:

$$x_{ij} = \mathcal{F}^{-1} \left[\left(p_A \otimes \hat{A}(x_{ij}) \right) * e^{-i * \mathcal{P}(x_i)(u,v)} \right], \quad (2)$$

where u and v represent coordinates in frequency domain. The generated image x_{ij} retains the semantic content of x_i with diversified dominant frequency components. This expands the detector focus across frequency bands. The Fo-Mixup augmentation is outlined in Algorithm 1.

Confidence Sampling. Augmented samples with high prediction entropy (i.e., low confidence) hinder generalization performance. To address this, we filter out unreliable samples from the pool of augmented data using Shannon entropy, defined as $E_{x_{ij}} = -\mathbf{p}(x_{ij}) \log(\mathbf{p}(x_{ij}))$, where x_{ij} is the augmented sample and $\mathbf{p}(x_{ij})$ denotes the model’s output probability for that sample. After calculating the entropy for each prediction, we rank the samples in ascending order and select the top λ percent of high-confidence samples.

3.2. Consistency Regularization

Our model consists of a feature extraction module with multiple stages, represented as $\Psi = \psi_M \circ \dots \circ \psi_2 \circ \psi_1$, where M denotes the number of stages in the network. Since each layer captures concepts at different abstraction levels [69],

from low-level textures to high-level semantics, we introduce a feature alignment module ω_i after each stage to capture regularized semantic abstractions progressively across the network. We express these as $F_i = \omega_i \circ \psi_i(x) \circ \dots \circ \psi_1(x)$, where $i \in 1, \dots, M$. The feature alignment module, which includes convolutional and residual blocks similar to those in ResNet [23], ensures that each auxiliary branch undergoes the same number of down-sampling operations as the backbone, supporting a fine-to-coarse feature transformation. With the feature alignment modules in place, the backbone is trained and its class posterior distributions for the predictions of the synthesized $(x^s, y) \in S$ and training images $(x^t, y) \in T$ are regularized as follows:

$$L_{cls} = \mathcal{L}_{CE}(\sigma(\alpha_f(x^s)), y) + \mathcal{L}_{CE}(\sigma(\alpha_f(x^t)), y) + D_{KL}(\sigma(\alpha_f(x^s); \tau), \sigma(\alpha_f(x^t); \tau)), \quad (3)$$

where $\mathcal{L}_{CE}$ indicates the standard cross-entropy (CE) loss function, D_{KL} denotes the Kullback-Leibler (KL) divergence, and $\alpha_f(x)$ is the final prediction of our model. Also, $\sigma(\alpha(x); \tau)$ denotes the Softmax operation with temperature τ . When $\tau = 1$, it would be a normal Softmax operation, expressed as $\sigma(\alpha_i(x))$.

Attention Consistency Regularization. With the synthesized forgery samples covering a wider range of dominant frequency components, we first regularize our detector using local consistency to identify consistent visual cues, regardless of dominant frequency variations. To achieve this, we compute CAMs [51, 72] on the feature map $F_M \in \mathbb{R}^{c \times h \times w}$ at the final stage of our backbone, where c , h , and w represent the number of channels, height, and width, respectively. CAMs offer valuable interpretability by highlighting the critical regions in an image that are relevant for detecting forgeries. The proposed Fo-Mixup augmentation

is designed to maintain the spatial location of tampered regions consistently between the input image x^t and the synthesized forgery x^s , preserving critical information for detection. By enforcing CAM consistency, we ensure that our forgery detector remains focused on the discriminative regions, making it less susceptible to domain-specific noise or artifacts that could otherwise mislead its predictions. For consistency supervision at the attention level, we calculate CAMs using a global average pooling (GAP) layer followed by a fully connected (FC) layer with weights $W \in \mathbb{R}^{c \times n}$, where n , the number of classes, is two. The FC layer is trained using the CE loss function, L_{CAM} , to distinguish between forgery and real samples. The CE loss is defined as $L_{CAM} = -\sum_{i=1}^n y_i \log(\hat{y}_i)$, where y_i is the true label and $\hat{y}_i$ is the predicted probability for class i . The classifier weights W are convolved with the feature map F_M to compute the CAMs as follows:

$$M = W^\top F_M, \quad (4)$$

where $M \in \mathbb{R}^{n \times h \times w}$. This process ensures that the CAMs can effectively highlight the most discriminative regions for the forgery classification. The CAMs are then normalized within the class region so that the highly valued features are solely assigned to the class region. For class-wise normalization, the CAMs are average pooled and clustered into different clusters using the K-means algorithm [39]. The first cluster (formulated as $Mask^{high}$) highlights the class region and the remaining ones would be discarded. Then, the obtained category region with $Mask^{high}$ is further normalized by instance normalization [55] to have a standard distribution. These operations are summarized as follows:

$$M^{cn} = (M \otimes Mask^{high})\gamma + \beta, \quad (5)$$

where M^{cn} and $Mask^{high}$ denote the class-wise normalized CAMs and the highlighted class region at the last layer of our backbone. γ and β are learnable affine parameters in the instance normalization. After all, we match the refined CAMs (i.e., M^{cn}) between the synthesized and original training domains by the Jensen-Shannon Divergence (JSD) as follows:

$$L_{att} = D_{JS}(\sigma(M^{cn}(x^s); \tau), \sigma(M^{cn}(x^t); \tau)), \quad (6)$$

where $M^{cn}(x^s)$ and $M^{cn}(x^t)$ are the class-wise normalized CAMs for the synthesized and training images.

Hyperspherical Consistency Regularization. Recent research demonstrates the effectiveness of hyperspherical latent spaces over Euclidean spaces in modeling the complex geometric structure of facial feature representations [22, 28, 32, 57]. To leverage this, we introduce a vMF classifier, implemented as a mixture model with two vMF distributions, to statistically model facial feature representations on the hyperspherical space for CR. To capture

a global view of the embedding space for consistent geometric learning, we leverage hierarchical feature maps from different stages of the backbone network to construct our vMF classifier. To achieve this, feature maps from these stages are concatenated and resized as $F_{cat} = \text{Conv}(\text{Concat}([F_1, \dots, F_M]))$, where Conv denotes the 1×1 convolutional layer that reduces the channel dimension of the concatenated feature maps. Given a normalized facial feature representation $\tilde{F}_{cat} = F_{cat} / \|F_{cat}\|_2$, the probability density function (PDF) for i -th class is formulated as follows:

$$p(\tilde{F}_{cat} | \kappa_i, \tilde{\mu}_i) = \mathcal{C}_d(\kappa_i) \exp(\kappa_i \cdot \tilde{F}_{cat} \tilde{\mu}_i^\top), \quad (7)$$

$$\mathcal{C}_d(\kappa_i) = \frac{\kappa_i^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} \cdot I_{\frac{d}{2}-1}(\kappa_i)}, \quad (8)$$

where $\tilde{F}_{cat} \in \mathbb{S}^{d-1}$ is the d -dimensional unit vector, and $\kappa_i \in \mathbb{R}_{\geq 0}$ and $\tilde{\mu}_i \in \mathbb{S}^{d-1}$ are trainable compactness parameter and orientation vector of the vMF classifier for i -th class. Also, $I_\beta(\kappa)$ is the modified Bessel function of the first kind at order β and $\mathcal{C}_d(\kappa)$ is a normalization constant. Given the PDF in Eq. 7, and the posterior probability, the vMF classifier is trained using CE loss as below:

$$L_{cls_sphere} = \mathcal{L}_{CE}(p(y_i | \tilde{F}_{cat})), \quad (9)$$

$$p(y_i | \tilde{F}_{cat}) = \frac{n_i \cdot p(\tilde{F}_{cat} | \kappa_i, \tilde{\mu}_i)}{\sum_{j=1}^2 n_j \cdot p(\tilde{F}_{cat} | \kappa_j, \tilde{\mu}_j)}, \quad (10)$$

where n_i indicates the cardinality of class i , and $p(y_i | \tilde{F}_{cat})$ denotes the probability of $\tilde{F}_{cat}$ being associated with class i . To enforce the geometric CR, the Distribution Matching Score (DMS) is formulated as follows:

$$\text{DMS}(\mathbf{F}_{cat}^s, \mathbf{F}_{cat}^t) = \frac{1}{1 + D_{KL}(p(\tilde{\mathbf{F}}_{cat}^s | \kappa_s, \tilde{\mu}_s), p(\tilde{\mathbf{F}}_{cat}^t | \kappa_t, \tilde{\mu}_t))}, \quad (11)$$

where $0 < \text{DMS} < 1$ represents the overlap between the synthesized S and training domains T . When two domains are highly overlapped, $\text{DMS} \rightarrow 1$. Finally, to enhance CR, we maximize the DMS between the synthesized S and training domains T in vMF optimization as below:

$$L_{sphere} = \mathbb{E}[1 - \text{DMS}(\mathbf{F}_{cat}^s, \mathbf{F}_{cat}^t)], \quad (12)$$

where $\mathbb{E}$ is the average function.

Overall Loss. The overall objective function in our training optimization boils down to five loss functions as follows:

$$L_{total} = L_{cls} + \eta L_{CAM} + \delta L_{att} + \mu L_{cls_sphere} + \rho L_{sphere}, \quad (13)$$

where η , δ , μ , and ρ are hyperparameters that balance the impact of different loss functions. Having trained our proposed model end-to-end, we remove all introduced auxiliary classifiers during testing, incurring no additional computational cost compared to the baseline backbone.

4. Experiments

4.1. Experimental settings

Forgery Detection Datasets. For simplicity and fair comparison, we pursue recent deepfake detection studies [8, 62, 64, 65] and train our model on the FaceForencis++ (FF++) dataset [47]. FF++ is the widely-used video dataset with 1,000 original videos and 4,000 fake videos which are generated by four face manipulation methods, including DeepFake (DF) [2], Face2Face (FF) [53], FaceSwap (FS) [3], and NeuralTextures (NT) [54]. In terms of quality, three compression levels are utilized in FF++ generation, which are raw quality, high quality (HQ), and low quality (LQ). The HQ level is adopted to sample 32 frames in each video. To gauge the generalization and robustness ability of our model, we conduct extensive experiments on FF++, Deepfake Detection Challenge (DFDC) [1], DeepFake Detection Challenge Preview (DFDCP) [1], Deepfake Detection (DFD) [14], Celeb-DF-v1 (CDFv1) [33], and Celeb-DF-v2 (CDFv2) [33] datasets.

Implementation detail. For preprocessing and training, we adhere to the configurations outlined in DeepFakeBench [64] to maintain a fair comparison, with an input size of 256×256 pixels. ResNet-34 [23] is chosen as our backbone, which is initialized with the pre-trained ImageNet weights. To apply OHM to the filtered images in Algorithm 1, our forgery detector is pre-trained on the FF++ dataset using the CE loss function for 30 epochs with a batch size of 32. In the CR for the forgery class, we incorporate the Fo-Mixup augmentation along with standard augmentations like contrast, pixelation, and the real class CR undergoes only standard augmentations. Our framework is trained for 50 epochs, with a batch size of 32, and an initial learning rate of $1e-4$. The hyperparameters used in Eq. 3 and 13, Fo-Mixup augmentation, and confidence sampling are set to $\tau = 4$, $\eta = 0.5$, $\delta = 0.1$, $\mu = 1$, $\rho = 0.1$, $k = 8$, $t = 3$, and $\lambda = 0.5$. Also, ξ in Eq. 1 is randomly sampled within $[0.0, 1.0]$. To benchmark our method, we follow the deepfake detection studies [8, 62, 64, 65] and adopt the frame-level area-under-the-curve (AUC), and Equal Error Rate (EER) metrics.

4.2. Comparison with previous forgery methods

Overall Performance Across Diverse Datasets. To explore the generalization capability of the proposed method, we train our model on the FF++ (HQ) dataset and gauge its generalization performance over other forgery datasets us-

Method	In-domain	Cross-domain					
	FF++	CDFv1	CDFv2	DFD	DFDCP	DFDC	C-Avg.
Xception [10]	96.4	77.9	73.7	81.6	73.7	70.8	75.54
Meso4 [4]	60.8	73.6	60.9	54.8	59.9	55.6	60.96
Capsule [40]	84.2	79.1	74.7	68.4	65.7	64.7	70.52
X-ray [31]	95.9	70.9	67.9	76.6	69.4	63.3	69.62
FFD [11]	96.2	78.4	74.4	80.2	74.3	70.3	75.52
F3Net [46]	96.4	77.7	73.5	79.8	73.5	70.2	74.94
SPSL [36]	96.1	81.5	76.5	81.2	74.1	70.4	76.74
SRM [38]	95.8	79.3	75.5	81.2	74.1	70.0	76.02
CORE [41]	96.4	78.0	74.3	80.2	73.4	70.5	75.28
RECCE [5]	96.2	76.8	73.2	81.2	74.2	71.3	75.34
SLADD [6]	96.9	80.2	74.0	80.9	75.3	71.7	76.42
IID [24]	97.4	75.8	76.9	79.3	76.2	69.5	75.54
UCF [63]	97.1	77.9	75.3	80.7	75.9	71.9	76.34
LSDA [46]	-	86.7	83.0	88.0	81.5	73.6	82.56
FreqDebias (Ours)	97.5	87.5	83.6	86.8	82.4	74.1	82.88

Table 1. Performance comparison in both in-domain and cross-domain settings. All detectors are trained on the FF++ (HQ) dataset [47], and evaluated on other datasets using the **frame-level** AUC metric. “C-Avg.” represents the average of all cross-domain results, with the best-performing values highlighted in bold.

ing the frame-level AUC metric. As reported in Table 1, it is observed that the SOTA approaches experience pronounced performance drops on the cross-domain evaluations. However, the FreqDebias framework mainly generalizes better to novel forgeries and achieves SOTA performance. More precisely, the FreqDebias framework yields gains of 0.8%, 0.6%, 0.9%, and 0.5% in terms of AUC metric compared to the runner-up methods in the CDFv1, CDFv2, DFDC, and DFDCP test sets, respectively. From this experiment, we can deduce that the proposed FreqDebias framework substantially improves the generalization of the forgery detector to unseen forgery attacks. Also, as reported in column FF++ in Table 1, we yield great performance in the in-domain setting, compared to SOTA studies.

Cross-Manipulation Evaluations. As new manipulation techniques continuously emerge, our method should generalize well to unseen forgeries. To assess this, we follow the training scheme from DCL [52] and perform a fine-grained cross-manipulation evaluation on the FF++ dataset. Specifically, our forgery detector is tested on individual manipulation techniques while being trained on data from other techniques within the same dataset. As represented in Table 2, our proposed method achieves generalization improvements across four unseen forgery types, with improvements ranging from 0.08% to 3.85%. This experiment demonstrates that our learned representation can effectively distinguish forgery artifacts from real ones in different forgeries. More experimental comparisons using the video-level AUC metric for both cross-domain and in-domain settings are provided in the **supplementary material**.

Robustness Evaluations. To evaluate the robustness of our method against diverse unseen corruptions, we follow the

Methods	Train	DF	F2F	FS	NT
GFF [38]	DF	99.87	76.89	47.21	72.88
DCL [52]		99.98	77.13	61.01	75.01
IID [24]		99.51	-	63.83	-
SFDG [59]		99.73	86.45	75.34	86.13
FreqDebias (Ours)		99.82	88.10	75.92	88.45
GFF [38]	F2F	89.23	99.10	61.30	64.77
DCL [52]		91.91	99.21	59.58	66.67
SFDG [59]		97.38	99.36	73.54	72.61
FreqDebias (Ours)		98.41	99.44	74.37	76.46
GFF [38]	FS	70.21	68.72	99.85	49.91
DCL [52]		74.80	69.75	99.90	52.60
IID [24]		75.39	-	99.73	-
SFDG [59]		81.71	77.30	99.53	60.89
FreqDebias (Ours)		83.76	78.93	99.78	63.48
GFF [38]	NT	88.49	49.81	74.31	98.77
DCL [52]		91.23	52.13	79.31	98.97
SFDG [59]		91.73	70.85	83.58	99.74
FreqDebias (Ours)		92.35	74.61	83.24	99.83

Table 2. Cross-manipulation performance comparison of the proposed FreqDebias framework with SOTA studies on the FF++ test set. The results are based on the **video-level** AUC metric.

Model	Saturation	Contrast	Block	Noise	Blur	Pixel	Avg
Face X-ray [31]	97.6	88.5	99.1	49.8	63.8	88.6	81.2
LipForensics [20]	99.9	99.6	87.4	73.8	96.1	95.6	92.1
RealForensics [21]	99.8	<u>99.6</u>	98.9	79.7	95.3	98.4	95.2
CADDMM [13]	99.6	99.8	99.8	87.4	99.0	98.8	97.4
FreqDebias (Ours)	99.6	99.8	<u>99.7</u>	89.2	<u>98.2</u>	99.1	97.6

Table 3. Robustness evaluation in terms of **video-level** AUC on FF++ dataset. **Bold** and underline indicate the top and second-best robustness, respectively. “Avg.” denotes the mean AUC score.

LipForensics [20] setup with uncompressed raw FF++ data. We exclude shared augmentations in our CR to ensure a fair comparison. As shown in Table 3, our method demonstrates outstanding robustness, outperforming RealForensics [21] and CADDMM [13] with an impressive 97.6% average AUC.

4.3. Ablation Study

In this subsection, we assess the contribution of different components in our method to the generalization performance. Experiments follow the cross-domain setting from [6, 36] using FF++ (HQ) as training and CDFv2 and DFDCP as the test sets. Ablation studies on the effectiveness of the hyperparameters in Eq. 13, as well as the impact of t , k , p_A , and the spectral analysis of the Fo-Mixup augmentation, are all provided in the **supplementary material**.

Effectiveness of Fo-Mixup Augmentation. To assess the impact of Fo-Mixup augmentation and confidence sampling, we deactivate them individually and together. In our experiments, the baseline model is the ResNet-34 network [23] which is trained without the proposed CR. As reported in Table 4, Fo-Mixup augmentation significantly enhances the generalization performance compared to the baseline FreqDebias-0, both with (FreqDebias-3) and without (FreqDebias-1) the proposed loss functions. In

Model	Components					CDFv2		DFDCP	
	DA	L_{att}	L_{sphere}	Fo-Mixup	CS	AUC	EER	AUC	EER
FreqDebias-0	✓					69.2	36.0	61.3	40.7
FreqDebias-1				✓		74.8	31.3	68.1	36.0
FreqDebias-2	✓	✓	✓			77.3	29.4	73.6	32.2
FreqDebias-3	✓	✓	✓	✓		81.7	25.5	80.1	27.7
FreqDebias-4	✓	✓	✓	✓	✓	83.6	23.9	82.4	26.1
FreqDebias-5	✓		✓	✓	✓	79.1	27.7	77.4	29.5
FreqDebias-6	✓	✓		✓	✓	80.4	26.5	78.2	28.9
Model	Components					CDFv2		DFDCP	
	DA	L_{att}	L_{sphere}	Fo-Mixup	CS	AM [61]	AS [61]	AUC	EER
FreqDebias-7	✓	✓	✓	✓	✓			78.5	28.2
FreqDebias-8	✓	✓	✓	✓	✓		✓	76.0	30.3
Model	Components					CDFv2		DFDCP	
	DA	L_{att}	L_{sphere}	Fo-Mixup	CS	λ		AUC	EER
FreqDebias-9	✓	✓	✓	✓	✓	0.80		82.1	25.2
FreqDebias-10	✓	✓	✓	✓	✓	0.65		82.9	24.5
FreqDebias-11	✓	✓	✓	✓	✓	0.50		83.6	23.9
FreqDebias-12	✓	✓	✓	✓	✓	0.35		82.4	24.9

Table 4. Ablation study on the Fo-Mixup augmentation, confidence sampling (CS), and standard data augmentation (DA). All detectors are trained on the FF++ (HQ) dataset [47], and evaluated on other datasets using the **frame-level** AUC metric.

FreqDebias-1, the model is trained with Fo-Mixup augmentation and standard CE loss function, highlighting the standalone effectiveness of Fo-Mixup augmentation. FreqDebias-3 demonstrates its impact when combined with the proposed loss functions. In FreqDebias-4, we evaluate confidence sampling by comparing it to FreqDebias-3, where all synthesized samples are used without sampling. The results demonstrate that confidence sampling enhances generalization by identifying the informative samples.

Fo-Mixup vs. Frequency Augmentations. To demonstrate the superiority of Fo-Mixup, we replace it with traditional frequency-based augmentations such as amplitude swap (AS) and amplitude mix (AM) from [61], as shown in Table 4. Comparative evaluations in FreqDebias-7 and FreqDebias-8 verify Fo-Mixup as an effective augmentation for detecting face forgeries. AM indiscriminately modifies all frequency components, neglecting key frequencies, while AS aggressively swaps amplitude spectrums. In contrast, Fo-Mixup targets dominant frequency components, enhancing the detector’s exposure to critical frequency variations.

Effectiveness of Consistency Regularization. In this experiment, we keep the Fo-Mixup component and explore the effect of the proposed attention and hyperspherical CRs. As shown in Table 4, our FreqDebias framework shows significant performance drops in AUC on the CDFv2 dataset, approximately 4.5% and 3.2%, when the L_{att} and L_{sphere} regularizations are deactivated (in FreqDebias-5 and FreqDebias-6, respectively). We do not ablate the effect of L_{cls_sphere} since removing L_{cls_sphere} from the total loss function no longer optimizes the vMF classifier with learnable κ and μ , which are also used in L_{sphere} .

Choice of λ value. We investigate the impact of different λ values, representing the percentage of selected samples, on generalization performance. We test λ values of 0.35,

Model	CDFv2		DFDCP	
	AUC	EER	AUC	EER
FreqDebias-13 (ResNet-50)	83.9	23.7	82.9	25.7
FreqDebias-14 (ConvNeXt)	85.1	22.6	83.4	25.3

Table 5. Ablation study on different backbones. All detectors are trained on the FF++ (HQ) dataset [47], and evaluated on other datasets using the **frame-level** AUC metric.

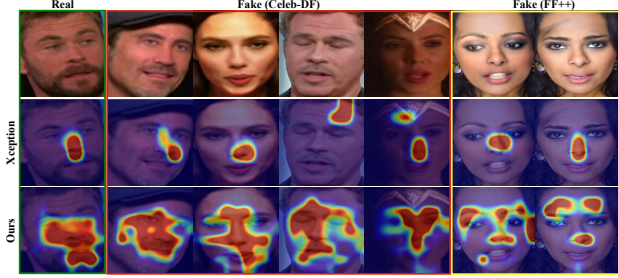


Figure 3. Visualization of the saliency maps for the Xception model [47] and the proposed FreqDebias framework.

0.50, 0.65, and 0.80 across FreqDebias-12 to FreqDebias-9 experiments in Table 4. Our findings reveal that while a moderate lambda value of 0.50 is beneficial, extreme values (too high or too low) cause performance degradation. With $\lambda = 0.80$ and $\lambda = 0.35$, the model shows a significant drop in AUC, indicating overfitting or insufficient augmentation. Therefore, we use $\lambda = 0.50$ in the remaining experiments to balance Fo-Mixup’s effectiveness without overfitting.

Effectiveness of Backbone. To evaluate the dependency of our performance on the ResNet-34 backbone, we substitute ResNet-34 with the ConvNeXt [37] and ResNet-50 backbones. The performance of our FreqDebias framework employing the ResNet-50 and ConvNeXt backbones (FreqDebias-13 and FreqDebias-14 in Table 5) surpasses the ResNet-34 backbone. This substantiates the adaptability of our framework for generalized forgery detection with different backbone choices.

4.4. Visualization Analysis

Analysis of Saliency Map. To localize and visualize critical facial regions in unseen forgeries, we apply CAM to our FreqDebias framework and the Xception model [47] trained on FF++. As seen in Fig. 3, Xception overfits to central facial regions, while FreqDebias effectively captures comprehensive regions, distinguishing between various forgery types and directing attention to respective artifacts, like eyebrow artifacts in FS forgeries. For instance, in the FS forgery (last column, Fig. 3), the FreqDebias framework focuses on artifacts in the eyebrow area.

Analysis of Feature Space. We employ t-SNE [56] to visualize the feature distribution learned by our FreqDebias framework. As shown in Fig. 4, compared to Xception [47]

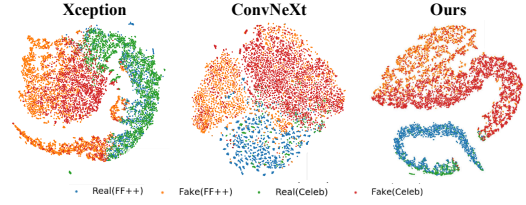


Figure 4. t-SNE illustration comparing the proposed FreqDebias framework with ConvNeXt [37] and Xception [47] models, all trained on the FF++ dataset.

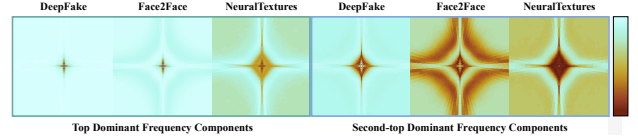


Figure 5. Heatmap of average dominant frequency components detected by Fo-Mixup using a standard deepfake detector.

and ConvNeXt [37], FreqDebias better clusters real samples with class-wise separation and fewer outliers. While it is not explicitly trained to classify forgery types, the directional constraints of the vMF distribution enable FreqDebias to capture underlying structures, leading to partial separation despite some overlap at finer scales. Moreover, FreqDebias shows superior discrimination between real and forgery samples with enhanced generalization.

Illustration of the Dominant Frequency Components.

Fig. 5 displays heatmaps of the average dominant frequency components heavily relied upon by a standard deepfake detector across 30,000 forgery images from FF++ [47]. For each forgery type, 10,000 images with similar identities are selected. In DeepFake [2] forgeries, the detector focuses solely on very low-frequency bands, ignoring mid and high-frequency bands. Face2Face [53] forgeries, however, also rely on mid-frequency bands. NeuralTextures [54] forgeries differ further, as the dominant frequency components also span high-frequency bands.

5. Conclusion

This paper presents FreqDebias, a novel frequency debiasing framework that mitigates spectral bias in deepfake detection to enhance generalization ability. FreqDebias includes two key strategies: (1) Forgery Mixup (Fo-Mixup), an augmentation technique that broadens the detector’s exposure to a diversified frequency spectrum, and (2) dual consistency regularization (CR), which explicitly regulates learning across frequency components via local and global consistency. Extensive experiments show that addressing spectral bias with FreqDebias achieves state-of-the-art generalization in in-domain and cross-domain settings.

References

- [1] Deepfake detection challenge. <https://www.kaggle.com/competitions/deepfake-detection-challenge/data>, 2019. 1, 6
- [2] Deepfakes github. <https://github.com/deepfakes/faceswap>, 2019. 6, 8
- [3] Faceswap. <https://github.com/MarekKowalski/FaceSwap>, 2019. 6
- [4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7. IEEE, 2018. 6
- [5] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022. 6
- [6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. 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*, pages 18710–18719, 2022. 1, 6, 7
- [7] Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. OST: Improving generalization of deepfake detection via one-shot test-time training. In *Advances in Neural Information Processing Systems*, pages 24597–24610, 2022. 1
- [8] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? In *Advances in Neural Information Processing Systems*, 2024. 6
- [9] Jikang Cheng, Ying Zhang, Qin Zou, Zhiyuan Yan, Chao Liang, Zhongyuan Wang, and Chen Li. Ed 4: Explicit data-level debiasing for deepfake detection. *arXiv preprint arXiv:2408.06779*, 2024. 1, 2
- [10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 6
- [11] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020. 6
- [12] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *European Conference on Computer Vision*, pages 18–35, 2022. 1
- [13] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 7
- [14] Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. <https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html>, 2019. 1, 6
- [15] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20270–20280, 2022. 1
- [16] Paul Gavrikov and Janis Keuper. Can biases in imagenet models explain generalization? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22184–22194, 2024. 2
- [17] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2
- [18] Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *European Conference on Computer Vision*, pages 596–613, 2022. 2
- [19] Ying Guo, Cheng Zhen, and Pengfei Yan. Controllable guide-space for generalizable face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20818–20827, 2023. 2
- [20] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 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*, pages 5039–5049, 2021. 7
- [21] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14950–14962, 2022. 7
- [22] Md Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, Liming Chen, et al. von Mises-Fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017. 2, 3, 5
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 6, 7
- [24] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiabin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2023. 1, 2, 6, 7
- [25] Jiaying Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. FSDR: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2
- [26] Mobarakol Islam and Ben Glocker. Frequency dropout: Feature-level regularization via randomized filtering. In *European Conference on Computer Vision*, pages 281–295. Springer, 2022. 2
- [27] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deepforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020. 1

- [28] Michael Kirchhof, Karsten Roth, Zeynep Akata, and Enkelejd Kasneci. A non-isotropic probabilistic take on proxy-based deep metric learning. In *European Conference on Computer Vision*, pages 435–454, 2022. 3, 5
- [29] Takumi Kobayashi. T-vMF similarity for regularizing intra-class feature distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6616–6625, 2021. 2, 3
- [30] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. 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*, pages 6458–6467, 2021. 1
- [31] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 2, 6, 7
- [32] Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical confidence learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15629–15637, 2021. 2, 3, 5
- [33] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deep-fake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 1, 6
- [34] Jiahao Liang, Huafeng Shi, and Weihong Deng. Exploring disentangled content information for face forgery detection. In *European Conference on Computer Vision*, pages 128–145, 2022. 1, 2
- [35] Manxi Lin, Nina Weng, Kamil Mikolaj, Zahra Bashir, Morten BS Svendsen, Martin G Tolsgaard, Anders N Christensen, and Aasa Feragen. Shortcut learning in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633. Springer, 2024. 2
- [36] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 772–781, 2021. 2, 6, 7
- [37] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 8
- [38] 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. 1, 2, 3, 6, 7
- [39] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of fifth Berkeley Symposium on Mathematical Statistics and Probability*, page 281–297, 1967. 3, 5
- [40] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2307–2311. IEEE, 2019. 6
- [41] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. CORE: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12–21, 2022. 2, 6
- [42] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 3
- [43] Hannah Pinson, Joerie Lenaerts, and Vincent Ginis. Linear cnns discover the statistical structure of the dataset using only the most dominant frequencies. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 2
- [44] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982. 3
- [45] Aahlad Puli, Nitish Joshi, Yoav Wald, He He, and Rajesh Ranganath. Nuisances via negativa: Adjusting for spurious correlations via data augmentation. *arXiv preprint arXiv:2210.01302*, 2022. 2
- [46] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103, 2020. 2, 6
- [47] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 6, 7, 8
- [48] Amirreza Shaban, Amir Rahimi, Thalaisyasingam Ajanthan, Byron Boots, and Richard Hartley. Few-shot weakly-supervised object detection via directional statistics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3920–3929, 2022. 2, 3
- [49] Kaede Shiohara and Toshihiko Yamasaki. Detecting deep-fakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 1, 2
- [50] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016. 3
- [51] Guolei Sun, Salman Khan, Wen Li, Hisham Cholakkal, Fahad Shahbaz Khan, and Luc Van Gool. Fixing localization errors to improve image classification. In *European Conference on Computer Vision*, pages 271–287, 2020. 4
- [52] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2316–2324, 2022. 6, 7

- [53] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 6, 8
- [54] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38(4):1–12, 2019. 6, 8
- [55] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017. 5
- [56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008. 8
- [57] Hualiang Wang, Siming Fu, Xiaoxuan He, Hangxiang Fang, Zuozhu Liu, and Haoji Hu. Towards calibrated hyper-sphere representation via distribution overlap coefficient for long-tailed learning. In *European Conference on Computer Vision*, pages 179–196, 2022. 2, 3, 5
- [58] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. Frequency shortcut learning in neural networks. In *NeurIPS Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 2
- [59] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. 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*, pages 7278–7287, 2023. 7
- [60] Jianqing Xu, Shen Li, Ailin Deng, Miao Xiong, Jiaying Wu, Jiaxiang Wu, Shouhong Ding, and Bryan Hooi. Probabilistic knowledge distillation of face ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3489–3498, 2023. 3
- [61] 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. 2, 3, 7
- [62] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023. 6
- [63] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023. 1, 2, 6
- [64] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, pages 4534–4565. Curran Associates, Inc., 2023. 6
- [65] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 6
- [66] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2
- [67] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022. 2
- [68] Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, and Pinar Duygulu. Hybridaugment++: Unified frequency spectra perturbations for model robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5718–5728, 2023. 2
- [69] Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Attributable visual similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7532–7541, 2022. 4
- [70] Lei Zhang, Ji-Fu Li, and Wei Wang. Semi-supervised domain generalization with known and unknown classes. In *Advances in Neural Information Processing Systems*, 2023. 2
- [71] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021. 1
- [72] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2, 4