Holcombe Department of ELECTRICAL AND COMPUTER ENGINEERING
Clemson University

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

Hossein Kashiani, Niloufar Alipour Talemi, Fatemeh Afghah

CVPR Nashville JUNE 11-15, 2025

Motivation

Why Do Deepfake Detectors Fail to Generalize?

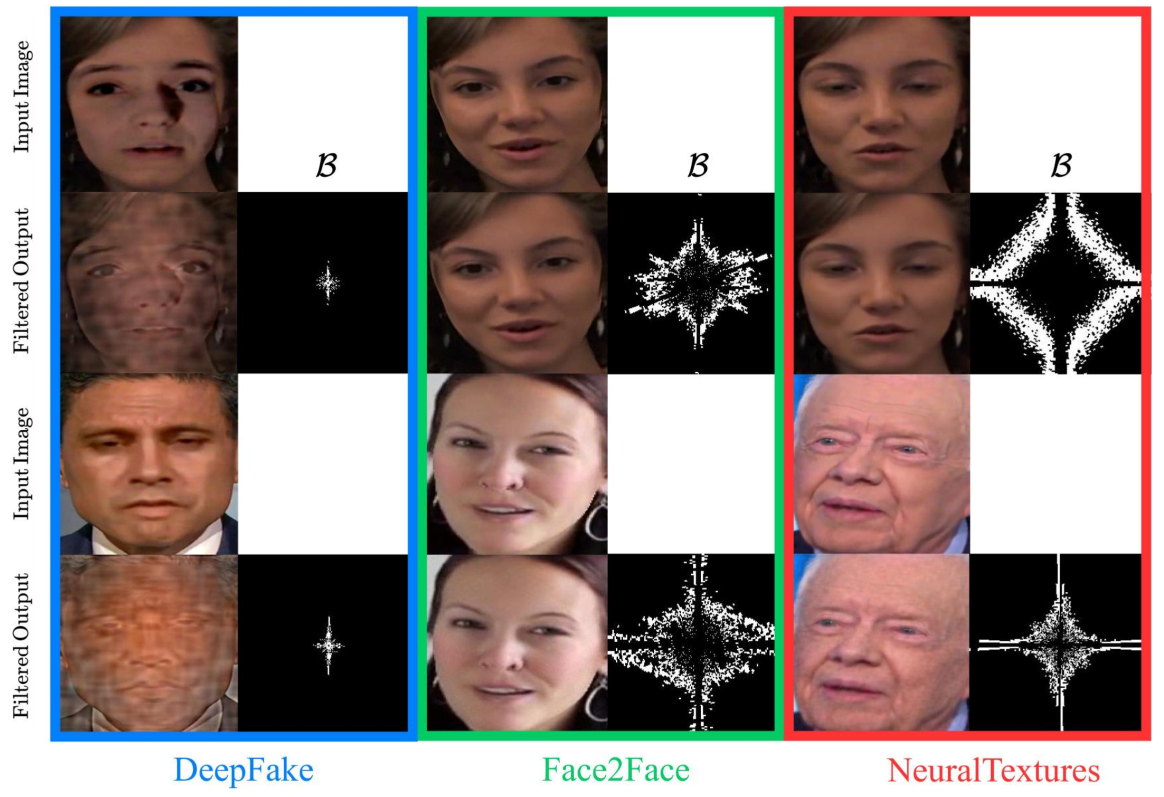
- Existing detectors exhibit poor cross-domain performance due to **model bias**.
- Detectors rely on **spurious correlations** such as identity, background, or structural artifacts.
- Prior works focus on **human-perceptible biases**, while this work investigates a form of model bias that is **imperceptible** to humans, known as **spectral bias** in the frequency domain.

Key Insights

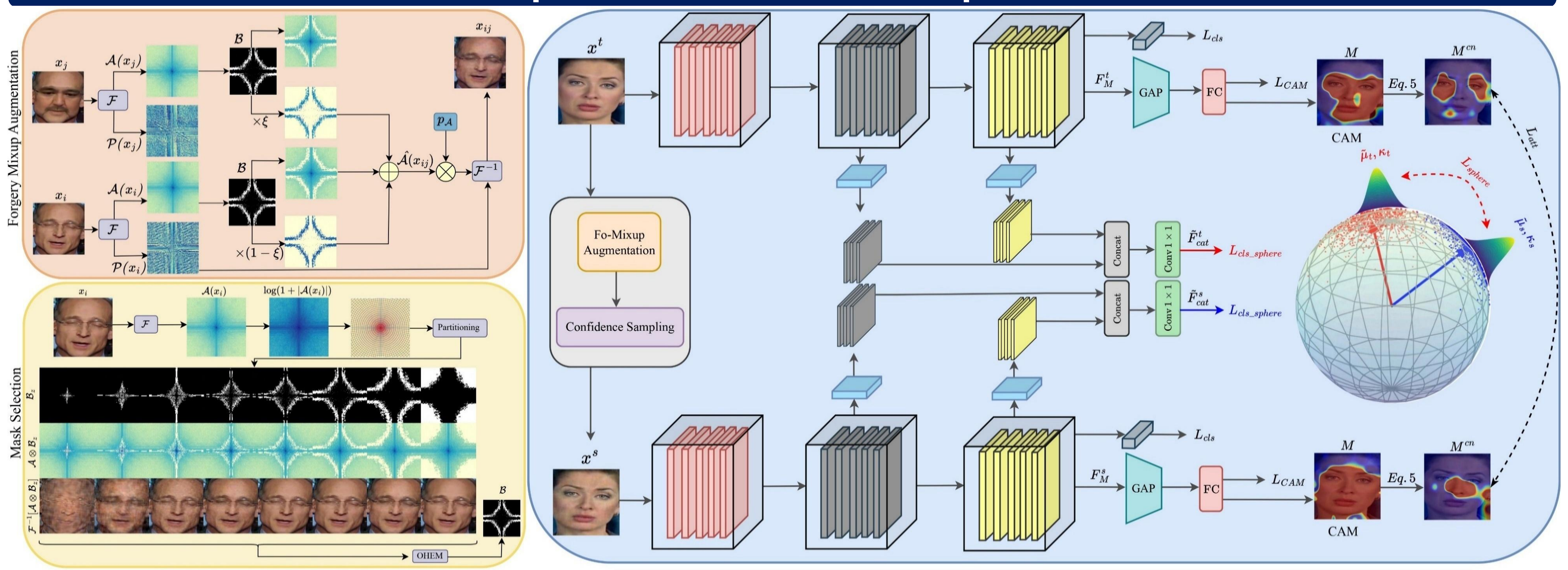
What is Spectral Bias?

- Detectors over-rely on **dominant frequency components**, which are specific to forgery types.
- These components are identified as frequency bands whose exclusion causes the largest increase in classification loss.
- Such reliance limits generalization to **unseen forgeries**.

Dominant frequency components overly relied upon by the vanilla deepfake detector.



Proposed Framework: FreqDebais



Forgery Mixup Augmentation:

- Identifies **dominant frequency components** and modulates the amplitude spectra within these components.
- Filters low-confidence augmented samples using Shannon entropy.

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

Dual Consistency Regularization:

- **Local Consistency**: Enforced via **Class Activation Maps (CAMs)** to maintain attention on discriminative regions.
- **Global Consistency:**

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

  - ✓ Model facial features on a **hyperspherical** embedding space using **von Mises-Fisher (vMF)** distribution.
  - ✓ Enforce domain alignment using the Distribution Matching Score: $\mathrm{DMS} = 1\Big/\left(1 + D_{\mathrm{KL}}\left(p(\tilde{\boldsymbol{F}}^s_{cat} \mid \kappa_s, \tilde{\mu}_s),\ p(\tilde{\boldsymbol{F}}^t_{cat} \mid \kappa_t, \tilde{\mu}_t)\right)\right)$

$$L_{sphere} = \mathbb{E}\left[1 - \mathrm{DMS}(\boldsymbol{F}^s_{cat}, \boldsymbol{F}^t_{cat})\right]$$

$$L_{total} = L_{cls} + \eta L_{CAM} + \delta L_{att} + \mu L_{cls\_sphere} + \rho L_{sphere}$$

Experimental Results

In-domain and Cross-domain Results

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 [65]	-	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**

Cross-manipulation Results

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**

Experiments

In-Domain and Cross-Domain Evaluations:

- **Trained on** FF++ (HQ); **Evaluated on** CDFv1, CDFv2, DFDC, DFDCP, DFD.

Cross-Manipulation Evaluations:

- Trained on one manipulation type of FF++ (e.g., DF) and tested on others.

Robustness Evaluations:

- Evaluated on six distortion types from LipForensics benchmark.

Different Backbones

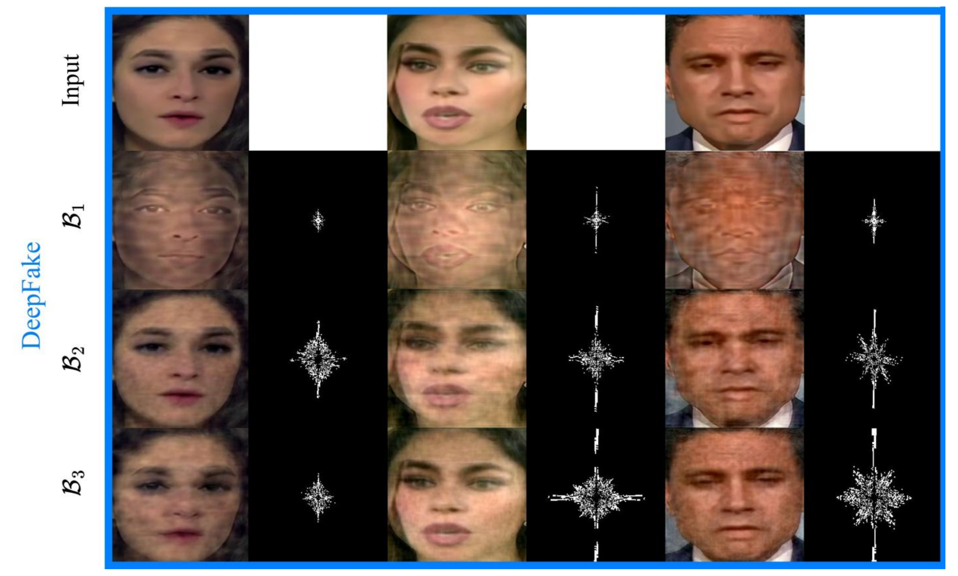
Model	CDFv2		DFDCP	
	AUC	EER	AUC	EER
ResNet-50	83.9	23.7	82.9	25.7
ConvNeXt	85.1	22.6	83.4	25.3

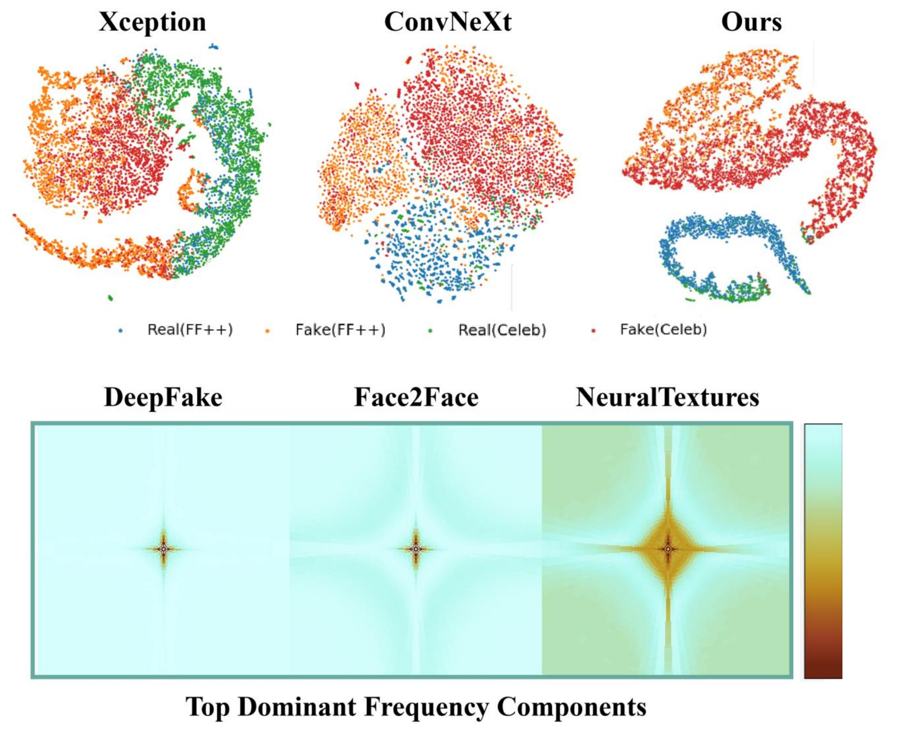
Robustness Results

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	99.6	98.9	79.7	95.3	98.4	95.2
CADDM [13]	99.6	**99.8**	**99.8**	87.4	**99.0**	98.8	97.4
FreqDebias (Ours)	99.6	**99.8**	99.7	**89.2**	98.2	**99.1**	**97.6**

Visualizations

- Standard detectors show forgery-specific dominant frequency reliance. Fo-Mixup targets this bias.





Conclusion

- **Spectral Bias**: We identify an **unexplored** form of model bias in deepfake detection.
- **Fo-Mixup:** We propose Fo-Mixup to broaden detector's exposure to a diversified frequency spectrum.
- **FreqDebias:** We propose FreqDebias, which first diversifies the frequency spectrum, and then enforces both local (CAMs) and global (vMF) consistency.
- **Experiments:** We demonstrate that FreqDebias significantly improves generalization across cross-domain and robustness settings.

Acknowledgment: This material is based upon work supported by the National Science Foundation under Grant Numbers CNS2232048, and CNS-2204445.