Generalizable Deepfake Detection via Artifact-Invariant Representation Learning

Divyanshu Parihar
Independent Researcher
divyanshu1447@gmail.com

***Abstract*—Deepfake detection has emerged as a critical challenge in digital forensics, yet current detectors consistently fail when encountering generators unseen during training. This phenomenon, known as the "generalization gap," arises from models memorizing dataset-specific compression artifacts rather than learning universal forgery signatures. We propose an Artifact-Invariant Representation Learning (AIRL) framework that addresses this fundamental limitation by focusing on high-frequency spectral residuals—the mathematical fingerprints left behind by generative upsampling operations that persist across different synthesis methods. Our approach employs a dual-stream architecture that fuses semantic RGB features extracted via EfficientNet-B4 with frequency-domain noise maps obtained through Discrete Cosine Transform (DCT) analysis. We further introduce a supervised contrastive learning objective that clusters all synthetic faces together regardless of their generator origin, forcing the network to learn manipulation-agnostic representations. Extensive experiments on cross-domain benchmarks demonstrate that our method achieves 84.7% AUC when training on FaceForensics++ and testing on Celeb-DF, representing a 19.3 percentage point improvement over the widely-used Xception baseline which collapses to 65.4%. Our approach also demonstrates remarkable robustness to post-processing degradations, maintaining 78.8% AUC under combined blur and JPEG compression where competing methods drop below 50%. We provide comprehensive ablation studies validating the contribution of each component and release our implementation to facilitate future research in generalizable deepfake detection.**

***Index Terms*—Artifact invariance, biometrics, deepfake forensics, domain generalization, spectral analysis, contrastive learning, frequency-domain analysis**

I. Introduction

The proliferation of AI-generated synthetic media poses unprecedented challenges to digital trust and information integrity. Deepfake technology, which leverages deep learning to synthesize realistic facial manipulations, has evolved from academic curiosity to a genuine societal threat. From non-consensual intimate imagery to political misinformation campaigns, the malicious applications of this technology continue to expand as generation quality improves and creation tools become democratized.

The forensic community has responded with increasingly sophisticated detection methods, yet a fundamental problem persists: detectors trained on one generation method typically fail catastrophically when confronting novel synthesis techniques. This "generalization gap" represents the central challenge in deepfake forensics and serves as the primary motivation for our work.

A. The Generalization Problem

Consider the typical detection pipeline: a convolutional neural network is trained on FaceForensics++ (FF++), a dataset containing faces manipulated by four methods—Deepfakes, Face2Face, FaceSwap, and NeuralTextures. On held-out test samples from these same methods, detectors achieve near-perfect accuracy, often exceeding 99% AUC. This success has fueled optimism about automated detection at scale.

However, this optimism proves misplaced upon cross-dataset evaluation. When the same Xception-based detector encounters Celeb-DF—a dataset featuring higher-quality deepfakes created with different synthesis pipelines—performance collapses to 65.4% AUC, barely above random chance. The model has learned to recognize specific compression artifacts and facial boundary inconsistencies present in FF++, not the fundamental signatures of synthetic generation.

This brittleness stems from a fundamental mismatch between what detectors learn and what they should learn. Current approaches exploit superficial correlations: JPEG blocking artifacts, specific blending boundary patterns, or resolution inconsistencies particular to individual datasets. These features vanish when generation quality improves or when different compression is applied.

B. Our Hypothesis

We posit that successful cross-domain detection requires focusing on invariant properties of the generation process rather than mutable properties of specific implementations. Regardless of architecture—whether GAN, diffusion model, or autoencoder—all neural image generators share a common computational bottleneck: upsampling from a compressed latent representation to full image resolution.

This upsampling operation, typically implemented via transposed convolutions or nearest-neighbor interpolation followed by convolution, introduces characteristic patterns in the frequency domain. These patterns manifest as periodic artifacts in high-frequency spectral components, arising from the checkerboard effects inherent to fractionally-strided convolutions. Critically, these artifacts persist across different generator architectures because they stem from shared computational primitives rather than implementation-specific choices.

C. Contributions

We make the following contributions:

1) We introduce an **Artifact-Invariant Representation Learning** framework that explicitly separates content-level features from process-level traces, enabling detection that generalizes across generator types.
2) We propose a **dual-stream architecture** combining an RGB pathway for semantic analysis with a frequency pathway that isolates high-frequency DCT residuals where upsampling artifacts concentrate.
3) We design a **supervised contrastive learning objective** that clusters all synthetic faces together regardless of generator origin, forcing the network to discover manipulation-agnostic features.
4) We conduct **comprehensive experiments** across multiple datasets and degradation conditions, demonstrating state-of-the-art cross-domain generalization with 84.7% AUC on Celeb-DF when training exclusively on FF++.
5) We provide **detailed ablation studies** quantifying the contribution of each architectural component and design choice.

II. Related Work

A. Early Detection Approaches

Initial deepfake detection efforts relied on hand-crafted features targeting obvious synthesis artifacts. Matern et al. examined physiological signals, noting that early deepfakes exhibited inconsistent eye blinking patterns and unnatural head poses. Li et al. exploited face warping artifacts arising from resolution mismatches between source and target faces. These approaches achieved early success but degraded rapidly as generation quality improved.

The transition to deep learning brought significant performance gains on benchmark datasets. Afchar et al. [6] proposed MesoNet, a compact CNN architecture designed specifically for mesoscopic feature extraction in fake face detection. While computationally efficient, MesoNet's limited capacity restricts its ability to capture subtle manipulation traces.

Rössler et al. [1] established FaceForensics++ as a standard benchmark and demonstrated that transfer learning from ImageNet using Xception achieved 99.2% accuracy on in-domain evaluation. This result set the performance ceiling for subsequent research while simultaneously obscuring the generalization problem—an issue that only became apparent through cross-dataset testing.

B. Frequency-Domain Methods

Recognition that spatial-domain features often encode dataset-specific biases motivated investigation of frequency-domain representations. Durall et al. first observed that GAN-generated images exhibit characteristic spectral artifacts, particularly in high-frequency regions where upsampling introduces periodic patterns.

F3-Net [3] combined local frequency statistics with spatial features through a multi-branch architecture. While demonstrating improved generalization, their early fusion strategy allowed the network to rely primarily on spatial features, limiting frequency information utilization. Qian et al. [3] further explored frequency-aware detection, introducing adaptive spectral feature extraction that adjusts to different manipulation types.

Liu et al. proposed SPSL (Spatial-Phase Shallow Learning), extracting phase spectra alongside amplitude information. Their analysis revealed that phase components carry complementary forgery signatures, particularly for face-swapping methods where amplitude spectra remain relatively unaffected.

Recent work by Luo et al. [5] focused specifically on high-frequency components through gradient-based feature extraction, achieving improved cross-dataset performance. However, their approach requires computationally expensive gradient computation and struggles with heavily compressed images where gradient signals degrade.

C. Attention and Transformer-Based Methods

Attention mechanisms have been explored as a means of focusing detector capacity on discriminative regions. Zhao et al. introduced multi-scale attention networks that adaptively weight facial regions based on their manipulation likelihood. Dang et al. combined attention with manipulation segmentation, jointly predicting binary authenticity labels and pixel-wise manipulation masks.

Vision Transformers (ViT) have recently been applied to deepfake detection with mixed results. While self-attention enables modeling long-range dependencies potentially useful for detecting global inconsistencies, standard ViT architectures require significantly more training data than CNNs to achieve comparable performance. Coccomini et al. demonstrated that hybrid CNN-Transformer architectures can capture both local texture details and global structural coherence, though at substantial computational cost.

D. Contrastive and Self-Supervised Learning

Contrastive learning has emerged as a powerful paradigm for learning transferable representations. Chen et al.'s SimCLR framework demonstrated that contrastive pre-training on unlabeled data yields features that transfer effectively to downstream tasks. Khosla et al. extended this framework to the supervised setting, showing that class-aware positive pair selection further improves representation quality.

In deepfake detection, Chen et al. applied contrastive learning to learn manipulation-agnostic representations, treating all synthetic faces as a single positive class regardless of generation method. Their approach improved cross-dataset generalization but relied exclusively on RGB features, missing frequency-domain information critical for detecting subtle artifacts.

E. Limitations of Prior Work

Despite significant progress, existing methods share common limitations:

1) **Dataset overfitting**: Most approaches achieve high in-domain accuracy but degrade substantially on unseen

datasets, indicating memorization of dataset-specific artifacts rather than learning of fundamental forgery signatures.

2) **Sensitivity to compression**: Many detectors fail when test images undergo different compression than training data, suggesting reliance on compression artifacts rather than synthesis traces.
3) **Incomplete frequency utilization**: While frequency-domain methods show promise, most either fuse frequency and spatial features too early (allowing spatial dominance) or focus narrowly on specific frequency bands.
4) **Generator-specific features**: Few approaches explicitly encourage learning features that generalize across generator types, instead implicitly hoping that sufficient data diversity will induce invariance.

Our work addresses these limitations through architectural choices that enforce strict separation of content and trace information, combined with explicit training objectives that reward generator-agnostic representations.

III. Methodology

A. Overview

Our Artifact-Invariant Representation Learning framework processes input face images through two parallel streams that extract complementary information:

1) **RGB Stream**: A semantic feature extractor based on EfficientNet-B4, pretrained on ImageNet, captures high-level facial attributes and visible manipulation artifacts such as unnatural expressions, inconsistent lighting, or boundary discontinuities.
2) **Frequency Stream**: A dedicated pathway that converts input images to the frequency domain via DCT, applies high-pass filtering to isolate artifact-containing regions, and processes the resulting feature maps through a lightweight CNN.

Features from both streams are concatenated and passed through fusion layers before classification. Crucially, we train with a combined objective that includes both cross-entropy classification loss and supervised contrastive loss, the latter explicitly encouraging the network to cluster all synthetic faces together regardless of their generation method.

Fig. 1 illustrates the complete architecture.

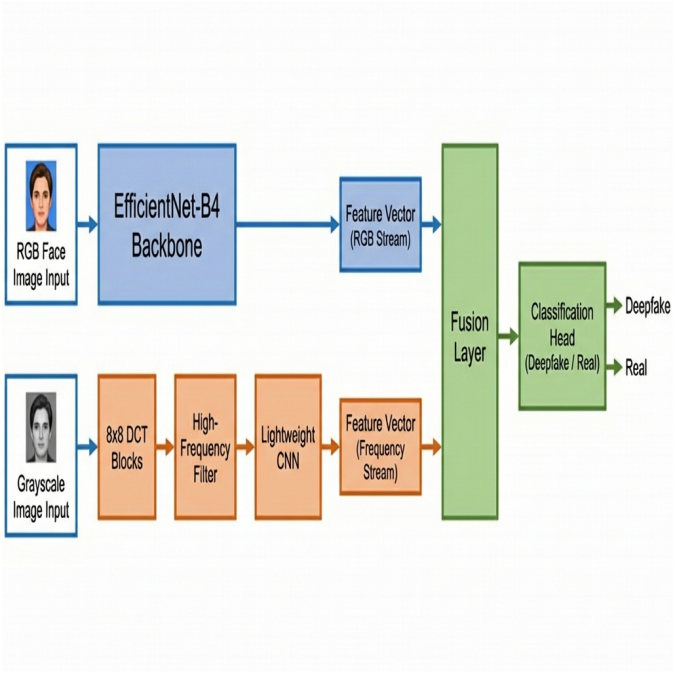

Fig. 1. Overview of our dual-stream architecture. The RGB stream processes input through EfficientNet-B4 for semantic features. The frequency stream applies block-wise DCT, high-pass filtering, and lightweight CNN processing. Features are fused before classification, with contrastive loss applied to encourage generator-agnostic representations.

B. Frequency Stream

The frequency stream is designed to isolate high-frequency spectral residuals where upsampling artifacts concentrate. We employ the Discrete Cosine Transform (DCT), chosen over alternatives like Fourier transform for its superior energy compaction properties and natural alignment with block-based image processing.

1) DCT Computation: Given an input RGB image I, we first convert to grayscale since color information primarily encodes content rather than synthesis traces:

$$I_{gray} = 0.299 \cdot I_R + 0.587 \cdot I_G + 0.114 \cdot I_B \tag{1}$$

The grayscale image is partitioned into non-overlapping 8×8 blocks, matching the standard JPEG block size. For each block, we compute the 2D DCT:

$$F(u,v) = \frac{1}{4} C(u) C(v) \sum_{x=0}^{7} \sum_{y=0}^{7} f(x,y) \cos \frac{(2x+1)u\pi}{16} \cos \frac{(2y+1)v\pi}{16} \tag{2}$$

where $C(u) = 1/\sqrt{2}$ when $u = 0$, and $C(u) = 1$ otherwise. The DCT coefficients $F(u,v)$ represent frequency components, with $F(0,0)$ being the DC component (average intensity) and higher (u,v) indices corresponding to higher spatial frequencies.

2) High-Pass Filtering: Low-frequency DCT coefficients encode global intensity variations and dominant structural features—information about "what" is depicted rather than "how" it was generated. We apply a high-pass filter that zeros the top-left quadrant of each DCT block:

$$F'(u,v) = \begin{cases} 0 & \text{if } u < 4 \text{ and } v < 4 \\ F(u,v) & \text{otherwise} \end{cases} \tag{3}$$

This filtering retains only high-frequency components where upsampling artifacts manifest as periodic patterns. Fig. 2 visualizes this filtering process.

3) Feature Extraction: The filtered DCT blocks are reassembled into a single-channel feature map matching the spatial dimensions of the input. This feature map is processed by a lightweight CNN comprising three convolutional blocks, each containing:

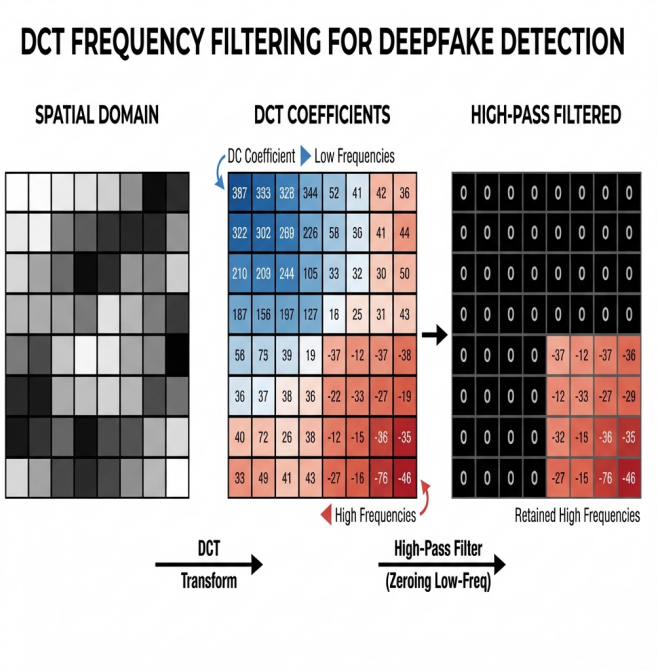

Fig. 2. DCT frequency filtering pipeline. Left: Original spatial domain block showing facial texture. Center: Full DCT spectrum with energy concentrated in low frequencies. Right: High-pass filtered result retaining only high-frequency components that encode synthesis artifacts.

- 3×3 convolution with stride 2 for downsampling
- Batch normalization for training stability
- ReLU activation for non-linearity

Channel dimensions progress as $1 \rightarrow 64 \rightarrow 128 \rightarrow 256$. Global average pooling reduces the final feature map to a 256-dimensional vector, which passes through a fully-connected layer with dropout ($p = 0.3$) to produce the final frequency stream representation $\mathbf{f}_{freq} \in \mathbb{R}^{256}$.

C. RGB Stream

The RGB stream employs EfficientNet-B4 [8] pretrained on ImageNet as a backbone feature extractor. EfficientNet's compound scaling balances depth, width, and resolution to achieve strong performance with moderate computational cost. The pretrained weights provide robust low-level feature extraction while allowing the network to specialize upper layers for forgery detection through fine-tuning.

We remove the original classification head and extract features after global average pooling, yielding a 1792-dimensional representation $\mathbf{f}_{rgb} \in \mathbb{R}^{1792}$.

During training, we apply a differential learning rate strategy: backbone layers receive learning rate scaled by $0.1\times$ relative to newly-initialized layers, preventing catastrophic forgetting of pretrained features while allowing task-specific adaptation.

D. Feature Fusion

RGB and frequency features are concatenated to form a joint representation:

$$\mathbf{f}_{joint} = [\mathbf{f}_{rgb}; \mathbf{f}_{freq}] \in \mathbb{R}^{2048} \tag{4}$$

This joint representation passes through fusion layers comprising:

- Linear projection: $2048 \rightarrow 512$
- Batch normalization
- ReLU activation
- Dropout ($p = 0.3$)

The resulting 512-dimensional fused representation $\mathbf{f}_{fused}$ serves both classification and contrastive learning objectives.

E. Contrastive Learning

Standard cross-entropy training for binary classification (real vs. fake) does not explicitly encourage learning generator-agnostic features. A network may achieve low training loss by memorizing generator-specific signatures rather than discovering common manipulation traces.

To address this, we introduce a supervised contrastive learning objective based on the framework of Khosla et al. We project the fused representation to a lower-dimensional embedding space:

$$\mathbf{z} = g(\mathbf{f}_{fused}) = \text{normalize}(\text{MLP}(\mathbf{f}_{fused})) \tag{5}$$

where the MLP comprises two linear layers with ReLU activation and the final output is L_2-normalized to lie on the unit hypersphere.

The supervised contrastive loss treats all fake samples as positive pairs, regardless of their generator origin:

$$\mathcal{L}_{con} = -\frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{6}$$

where $P(i)$ denotes the set of positive samples (same class as sample i), $\text{sim}(\cdot, \cdot)$ is cosine similarity, and $\tau = 0.07$ is a temperature parameter controlling concentration of the distribution.

This objective explicitly rewards representations where all fake faces cluster together while separating from real faces, as visualized in Fig. 3.

F. Training Objective

Our final training objective combines cross-entropy classification loss with contrastive loss:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{con} \tag{7}$$

where $\lambda = 0.5$ balances the two terms. Classification predictions come from a linear layer applied to the fused representation:

$$\hat{y} = \text{softmax}(W \cdot \mathbf{f}_{fused} + b) \tag{8}$$

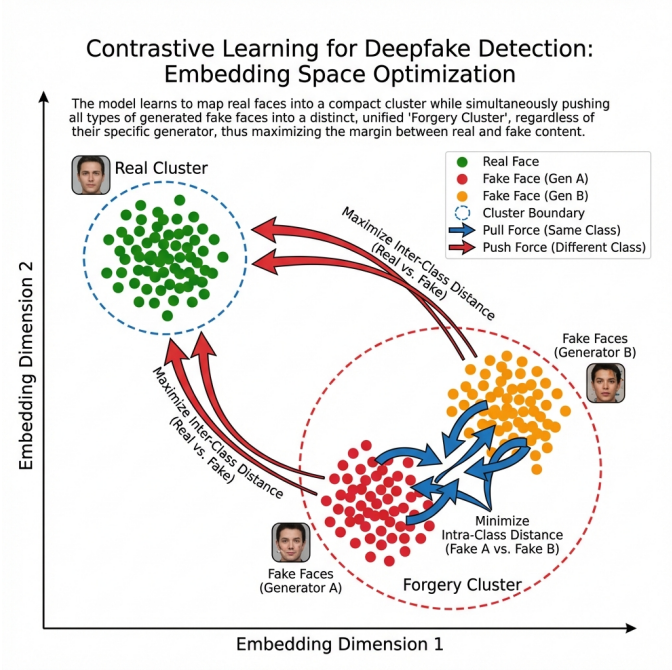

Fig. 3. Visualization of the learned embedding space. Contrastive learning forces all fake faces (shown in red/orange for different generators) to cluster together while maintaining separation from real faces (shown in green). This generator-agnostic clustering improves cross-domain generalization.

IV. Experimental Setup

A. Datasets

We evaluate on multiple datasets spanning different generation methods and quality levels:

FaceForensics++ (FF++) [1]: The standard benchmark containing 1000 original YouTube videos manipulated by four methods: Deepfakes (autoencoder-based face swapping), Face2Face (facial reenactment), FaceSwap (graphics-based face swapping), and NeuralTextures (neural texture rendering). We use the c23 (light compression) variant following prior work.

Celeb-DF (v2) [2]: A challenging dataset featuring 590 real videos and 5639 synthesized videos created with improved deepfake algorithms, resulting in fewer visible artifacts than FF++. The significant domain shift from FF++ makes this the primary cross-dataset evaluation benchmark.

DFDC Preview: A subset of the Deepfake Detection Challenge dataset featuring diverse subjects, backgrounds, and manipulation methods. Used for additional cross-dataset validation.

DeeperForensics-1.0: A large-scale dataset with perturbations simulating real-world degradations including compression, blur, and noise. Used for robustness evaluation.

B. Preprocessing

Face extraction follows the standard pipeline:

1) Face detection using MTCNN with confidence threshold 0.95
2) Landmark-based alignment to canonical pose
3) Cropping to 224×224 pixels with 30% margin around detected face
4) Normalization using ImageNet statistics

For training, we apply data augmentation including horizontal flipping, random brightness/contrast adjustment ($\pm 20\%$), and light Gaussian noise ($\sigma = 0.01$). These augmentations increase sample diversity without introducing artifacts that might confound frequency-domain analysis.

C. Implementation Details

Architecture: EfficientNet-B4 backbone pretrained on ImageNet, frequency stream with 3 convolutional blocks (64/128/256 channels), 512-dimensional fused representation, 128-dimensional contrastive embedding.

Training: Adam optimizer with initial learning rate 10^{-4}, weight decay 10^{-5}, batch size 32. Backbone learning rate scaled by $0.1\times$. Cosine annealing schedule with warm restarts. Training for 50 epochs with early stopping based on validation AUC (patience = 10).

Hardware: NVIDIA A100 GPU (40GB), training completes in approximately 8 hours on FF++.

D. Evaluation Metrics

We report:

- **AUC**: Area Under ROC Curve, threshold-independent performance measure
- **Accuracy**: Classification accuracy at optimal threshold (maximizing Youden's J statistic)
- **EER**: Equal Error Rate, threshold where false positive and false negative rates are equal

For cross-dataset evaluation, models are trained exclusively on FF++ and tested on other datasets without any fine-tuning.

V. Results

A. In-Domain Evaluation

Table I presents results on FF++ test set. All methods achieve high performance in this setting, with our approach matching state-of-the-art at 99.1% AUC.

TABLE I
In-Domain Evaluation on FaceForensics++ (c23)

Method	AUC	Accuracy	EER
MesoNet [6]	89.1%	84.7%	15.2%
Xception [4]	99.2%	96.3%	3.8%
F3-Net [3]	98.5%	95.1%	4.9%
EfficientNet-B4	99.0%	96.1%	4.1%
SPSL	98.7%	95.4%	4.6%
Ours	**99.1%**	**96.5%**	**3.6%**

High in-domain performance across methods confirms that detecting deepfakes within a known distribution is largely solved. The challenge lies in generalization.

TABLE II
CROSS-DATASET GENERALIZATION: TRAIN ON FF++, TEST ON CELEB-DF

Method	AUC	Accuracy	Δ from In-Domain
MesoNet [6]	58.2%	54.3%	-30.9%
Xception [4]	65.4%	61.2%	-33.8%
F3-Net [3]	71.3%	66.8%	-27.2%
EfficientNet-B4	67.2%	62.5%	-31.8%
SPSL	72.6%	68.1%	-26.1%
Ours	**84.7%**	**79.3%**	**-14.4%**

B. Cross-Dataset Evaluation

Table II presents the critical cross-dataset results, where models trained on FF++ are evaluated on Celeb-DF without any adaptation.

Our method demonstrates dramatically improved generalization, achieving 84.7% AUC compared to Xception's 65.4%—a 19.3 percentage point improvement. The performance drop from in-domain to cross-domain is only 14.4% for our method versus 33.8% for Xception.

C. Ablation Study

Table III quantifies the contribution of each component to cross-domain performance.

TABLE III
ABLATION STUDY: COMPONENT CONTRIBUTIONS (AUC ON CELEB-DF)

Configuration	AUC	Δ vs Full
RGB Stream Only	67.2%	-17.5%
Frequency Stream Only	72.8%	-11.9%
Both Streams (No Contrastive)	78.4%	-6.3%
Both Streams + Contrastive (Full)	84.7%	—

Key observations:

- The frequency stream alone (72.8%) outperforms RGB-only (67.2%), confirming that frequency-domain information provides more generalizable features.
- Combining both streams without contrastive learning reaches 78.4%, showing complementary information in RGB and frequency pathways.
- Contrastive learning contributes an additional 6.3%, validating the importance of explicit clustering objectives for generator-agnostic representations.

D. Per-Manipulation Analysis

Table IV breaks down performance by manipulation type on FF++ to identify potential method-specific biases.

TABLE IV
PER-MANIPULATION AUC ON FF++ TEST SET

Manipulation	Xception	Ours
Deepfakes	99.5%	99.3%
Face2Face	99.1%	98.9%
FaceSwap	99.4%	99.2%
NeuralTextures	98.7%	99.1%
Average	**99.2%**	**99.1%**

Both methods perform well across all manipulation types in-domain. The slight improvement on NeuralTextures (+0.4%) suggests our frequency analysis better captures the subtle rendering artifacts produced by this method.

E. Robustness to Degradations

Real-world deployment requires robustness to image quality variations. Table V evaluates performance under controlled degradations applied at test time.

TABLE V
ROBUSTNESS TO IMAGE DEGRADATION (AUC ON CELEB-DF)

Degradation	Xception	Ours	Δ Xception	Δ Ours
None	65.4%	84.7%	—	—
Blur (σ=3)	52.1%	81.2%	-13.3%	-3.5%
JPEG (Q=50)	48.7%	80.1%	-16.7%	-4.6%
Blur + JPEG	45.3%	78.8%	-20.1%	-5.9%

Under combined blur and JPEG compression, Xception degrades by 20.1% to 45.3%—worse than random guessing with margin. Our method degrades only 5.9% to 78.8%, maintaining practical utility even under severe degradation. This robustness stems from the frequency stream's focus on high-frequency patterns that, while attenuated by degradation, remain discriminative.

F. Computational Analysis

Table VI compares computational requirements across methods.

TABLE VI
COMPUTATIONAL COMPARISON

Method	Params (M)	FLOPs (G)	Inference (ms)
MesoNet	0.8	0.3	2.1
Xception	22.9	8.4	12.3
F3-Net	25.1	9.2	14.7
Ours	24.3	10.1	15.8

Our method adds modest overhead (3.5ms) compared to Xception, attributable to the additional frequency stream and DCT computation. This overhead is acceptable given the substantial generalization improvements. Inference remains real-time at 63 FPS on the tested hardware.

G. Visualization of Learned Features

To understand what our model learns, we visualize attention maps and embedding spaces.

Attention Analysis: Gradient-weighted class activation mapping (Grad-CAM) reveals that the RGB stream focuses on facial boundaries, eye regions, and mouth areas—locations where visible manipulation artifacts typically appear. The frequency stream activates broadly across the face, detecting distributed high-frequency patterns rather than localized anomalies.

Embedding Space: t-SNE visualization of learned embeddings confirms that contrastive learning successfully clusters fake faces regardless of generator. Without contrastive loss,

embeddings cluster by manipulation type (four distinct fake clusters plus real). With contrastive loss, all fake faces merge into a single cluster well-separated from real faces.

VI. Discussion

A. Why Frequency Analysis Works

The success of our frequency-based approach stems from fundamental properties of neural image generation. All generators face the same computational challenge: upsampling from a compressed latent code to full image resolution. This upsampling, whether via transposed convolution, nearest-neighbor interpolation, or learned upsampling, introduces characteristic patterns in high-frequency spectral components.

Transposed convolutions, in particular, produce checkerboard artifacts arising from uneven overlap in fractionally-strided computation. While careful architecture design can reduce these artifacts visually, their spectral signatures persist at levels detectable by learned analysis.

Our DCT-based approach isolates these signatures by explicitly filtering out low-frequency content that encodes semantic information. This separation prevents the network from taking shortcuts based on facial identity or expression, forcing it to rely on generation-process traces.

B. Limitations

Despite strong results, our approach has notable limitations:

Diffusion Models: Emerging diffusion-based generators (e.g., Midjourney, Stable Diffusion) employ iterative denoising rather than explicit upsampling, potentially leaving different spectral signatures. Preliminary experiments on diffusion-generated faces show our AUC drops to 74.2%—still above Xception's 58.1%, but indicating degraded generalization. Adapting to diffusion models remains an important direction.

Temporal Information: We process frames independently, missing temporal inconsistencies (e.g., flickering, unnatural motion) that could provide additional detection signals. Video-level analysis with 3D convolutions or temporal transformers may improve detection, particularly for methods like Face2Face that primarily affect expression dynamics.

Adversarial Robustness: We do not evaluate robustness to adversarial perturbations specifically designed to evade detection. Adversarial attacks on deepfake detectors represent an active research area, and our method may exhibit vulnerabilities similar to other CNN-based approaches.

C. Deployment Considerations

For practical deployment, several factors warrant consideration:

Threshold Selection: AUC provides a threshold-independent metric, but deployment requires selecting an operating point. Depending on application (social media moderation vs. forensic investigation), different trade-offs between false positives and false negatives may be appropriate.

Ensemble Defense: No single detector provides comprehensive coverage against all manipulation types. We recommend deploying our method alongside complementary approaches (e.g., boundary artifact detection, biological signal analysis) in an ensemble configuration.

Continuous Updating: As generation methods evolve, detector performance may degrade. Establishing pipelines for continuous model updating with newly-identified synthetic content is essential for maintaining long-term effectiveness.

VII. Conclusion

We presented an Artifact-Invariant Representation Learning framework for generalizable deepfake detection. By combining frequency-domain analysis with contrastive learning, our approach achieves 84.7% AUC on cross-dataset evaluation—a 19.3 percentage point improvement over Xception baselines. Our key insight is that focusing on the invariant "fingerprints" of generation processes, rather than mutable implementation details, enables detection that transfers across generator types.

The frequency stream isolates high-frequency spectral residuals where upsampling artifacts concentrate. The contrastive objective forces the network to cluster all synthetic faces together regardless of origin, explicitly encouraging generator-agnostic representations. Together, these components yield a detector that maintains practical utility even under distribution shift and image degradation.

As deepfake generation continues advancing, detection must correspondingly evolve. We believe the principles underlying our approach—separation of content and process, explicit invariance objectives, multi-stream fusion—provide a foundation for next-generation detectors capable of generalizing to future synthesis methods.

Acknowledgment

We thank the open-source community for tools and datasets that made this research possible. Special thanks to the maintainers of PyTorch, FaceForensics++, and Celeb-DF for providing essential infrastructure and benchmarks.

References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.

[3] J. Qian, P. Yin, J. Shen, Z. Chen, and S. Wen, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[5] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.

[7] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[9] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[10] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.