

SpecXNet: A Dual-Domain Convolutional Network for Robust Deepfake Detection

Inzamamul Alam
Department of Computer Science and
Engineering
Sungkyunkwan University
Suwon, Republic of Korea
inzi15@g.skku.edu

Md Tanvir Islam
Department of Computer Science and
Engineering
Sungkyunkwan University
Suwon, Republic of Korea
tanvirnwu@g.skku.edu

Simon S. Woo*
Department of Computer Science and
Engineering
Sungkyunkwan University
Suwon, Republic of Korea
swoo@g.skku.edu

ABSTRACT

The increasing realism of content generated by GANs and diffusion models has made deepfake detection significantly more challenging. Existing approaches often focus solely on spatial or frequency-domain features, limiting their generalization to unseen manipulations. We propose the Spectral Cross-Attentional Network (SpecXNet), a dual-domain architecture for robust deepfake detection. The core **Dual-Domain Feature Coupler (DDFC)** decomposes features into a local spatial branch for capturing texture-level anomalies and a global spectral branch that employs Fast Fourier Transform to model periodic inconsistencies. This dual-domain formulation allows SpecXNet to jointly exploit localized detail and global structural coherence, which are critical for distinguishing authentic from manipulated images. We also introduce the **Dual Fourier Attention (DFA)** module, which dynamically fuses spatial and spectral features in a content-aware manner. Built atop a modified XceptionNet backbone, we embed the DDFC and DFA modules within a separable convolution block. Extensive experiments on multiple deepfake benchmarks show that SpecXNet achieves state-of-the-art accuracy, particularly under cross-dataset and unseen manipulation scenarios, while maintaining real-time feasibility. Our results highlight the effectiveness of unified spatial-spectral learning for robust and generalizable deepfake detection. To ensure reproducibility, we release the full code on [GitHub](#).

CCS CONCEPTS

- Computing methodologies → Computer vision tasks.

KEYWORDS

Deepfake Detection, Dual-Domain Learning, Fake Image Classification, Frequency Domain, Security

1 INTRODUCTION

The proliferation of generative AI techniques such as Generative Adversarial Networks (GANs) [33] and Diffusion [56, 59] models has made the creation of highly realistic synthetic images both accessible and efficient. While these technologies offer creative potential, they also introduce significant societal risks, including misinformation, identity fraud, and political manipulation [48]. This has led to an urgent demand for robust and generalizable deepfake detection systems capable of operating across diverse image domains and generative models.

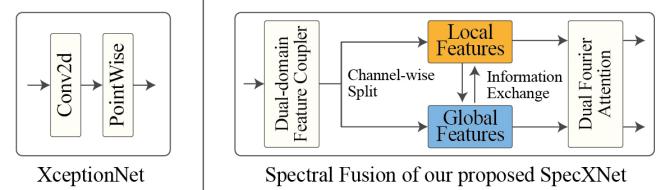


Figure 1: High-level architectural comparison between the original XceptionNet and our proposed modification. While XceptionNet utilizes standard depthwise separable convolutions (Conv2d + PointWise), our enhanced version decomposes each block into dual pathways: a local spatial branch using conventional convolutions and a global spectral branch powered by frequency-aware operations. These branches are then adaptively fused using the proposed Dual Fourier Attention (DFA) module to jointly exploit fine-grained local details and holistic spectral cues.

A range of deep learning methodologies has been proposed [4, 27, 39, 49, 52, 60, 61, 65, 67, 68, 68, 71] to address the deepfake detection problem, including preprocessing techniques designed to extract power spectral features for classification [15, 66]. Despite their initial success, these approaches often exhibit reduced effectiveness when applied to content generated by more recent and sophisticated generative models [2, 20, 55, 74]. In parallel, various forensic analysis techniques [16] have been explored; however, many existing fake image detection frameworks still lack the architectural robustness and accuracy necessary to reliably identify the latest AI-generated forgeries.

Recent state-of-the-art (SOTA) methods [10, 14, 50, 63, 64] have explored various avenues to address this challenge. Deepfake detection is commonly approached as a supervised classification task, where neural networks are trained to distinguish between genuine and altered visual content [10, 28, 46, 50]. However, a major limitation lies in their generalization capability of the models trained solely on specific types of synthetic images often struggle to accurately detect previously unseen manipulation techniques [28, 36]. Vision-language models (VLMs) like CLIP [54] have demonstrated impressive zero-shot capabilities for fake image detection by leveraging large-scale visual and textual representations. Adaptation strategies such as prompt tuning and adapter networks have enhanced CLIP's performance and generalizability across GAN and diffusion-based datasets [38, 72]. However, these approaches remain dependent on large-scale training and inference resources,

*Corresponding author. Email: swoo@g.skku.edu (Simon S. Woo)

often overlooking the energy efficiency and temporal dynamics crucial for real-time or edge deployment scenarios.

Simultaneously, frequency-based methods like UGAD [3] have attempted to harness unique spectral signatures of AI-generated images through Radial Integral Operations and Spatial Fourier Extraction. These techniques effectively distinguish real from fake images using frequency-domain fingerprints, achieving competitive accuracy across a variety of generative methods. However, such systems also rely on deep convolutional backbones like ResNet152 [24], which can be computationally expensive and may not fully exploit temporal information latent in the input data stream. Moreover, most existing methods operate in either the spatial or frequency domain, treating the two sources of evidence independently. This separation often causes subtle manipulation traces that are only evident in one domain to be overlooked.

To address these issues, we present SpecXNet, a deepfake detection framework that enhances the XceptionNet backbone by integrating spatial and spectral representations within a unified convolutional architecture. Unlike the original XceptionNet, SpecXNet introduces a dual-branch design: a local spatial branch that captures fine-grained textures via standard convolutions, and a global spectral branch that applies two-dimensional Fast Fourier Transform to model long-range frequency patterns. These complementary branches are fused through a novel Dual Fourier Attention (DFA) mechanism, which generates channel-wise attention maps to enable reciprocal modulation and content-adaptive fusion. This design promotes synergistic learning of spatial textures and spectral anomalies, significantly improving the model's ability to detect subtle traces of manipulation. SpecXNet embeds this dual-domain architecture within modified depthwise separable convolution blocks, referred to as Dual-Domain Feature Couplers (DDFC). As illustrated in Figure 1, these enhanced blocks offer a lightweight and scalable solution while delivering superior expressive power compared to the vanilla Xception architecture.

- Firstly, we propose a Dual-Domain Feature Coupler (DDFC) deepfake detection framework that jointly models spatial detail and spectral context through a bifurcated convolutional architecture. This enables the learning of robust multi-scale representations for better manipulation detection.
- Secondly, we introduce a novel Dual Fourier Attention (DFA) module that adaptively fuses local and global features using cross-domain modulation and attention-weighted integration, enhancing the semantic alignment of spatial and frequency cues.
- We demonstrate the effectiveness and generalization ability of our proposed SpecXNet through extensive evaluations of diverse generative models and benchmark datasets, showing consistent performance improvements over existing spatial and spectrum-based approaches.

2 RELATED WORKS

A wide range of approaches have been proposed in recent years to address the challenge of deepfake detection [4, 5, 11–13, 23, 27–30, 35, 37, 39, 40, 52, 60, 65, 67, 68]. Traditional supervised learning methods have been widely adopted, where models are trained to differentiate between authentic and manipulated content. For instance,

Wu et al. [64] adopted a conventional classification strategy, while Cozzolino et al. [17] introduced spectral preprocessing techniques that extract power spectrum features to enhance classification performance. Language-guided approaches have also emerged, with Radford et al. [54] leveraging perceptual learning guided by text supervision, and Wang et al. [64] presenting the DIRE model designed explicitly for diffusion-generated image detection. Similarly, Cozzolino et al. [17] proposed a GAN detection method based on a ResNet152 backbone. In another line of work, Zhang et al. [69] utilized Discrete Cosine Transform (DCT) to analyze power spectrum properties, though their evaluation was limited to a single diffusion model. Jeong et al. [31] employed spectral analysis by training generator-discriminator pairs and analyzing the resultant power spectrum distribution. In parallel, Radial Integral Operation (RIO) has been used to accumulate power spectrum density across radii, with findings indicating that deepfake images tend to exhibit consistent spectral distributions. This observation can be exploited for classification. Beyond spectral techniques, other forensic strategies have been explored. Corvi et al. [16] investigated noise print-based camera fingerprinting, while Mandelli et al. [45] focused on distinguishing real and manipulated Western blot images, albeit with limited generality. Ma et al. [42] applied a combination of statistical analysis and neural networks to detect synthesis artifacts in generative models. Unlike previous methods, our model unifies spatial and spectral analysis through a dual-branch architecture with adaptive attention, enabling robust and generalizable deepfake detection across both GAN- and diffusion-based manipulations.

3 PROPOSED METHOD: SPECXNET

We introduce the Spectral Cross-Attentional Network (SpecXNet), a convolutional neural network (CNN) architecture that jointly models local spatial and global spectral representations. As illustrated by the **cyan block** in Figure 2, the core design of SpecXNet lies in the explicit partitioning of input features into two complementary branches: one operating in the spatial domain and the other in the frequency domain. This separation enables the network to simultaneously capture fine-grained local patterns and broad contextual cues by applying standard convolutions to spatial features and Fast Fourier Transform to spectral features. The dual-domain architecture significantly improves the model's capacity to detect diverse manipulation artifacts across multiple scales and content types.

3.1 Dual-Domain Feature Coupler (DDFC)

Given an input feature tensor $X \in \mathbb{R}^{C \times H \times W}$, our proposed DDFC initiates a systematic decomposition of X to effectively leverage distinct yet complementary representations. As depicted by the **violet block** in Figure 2, we strategically partition the input into two specialized subsets: local spatial features X_l and global spectral features X_g . Formally, this decomposition is defined as follows:

$$(X_l, X_g) = \mathcal{D}(X; \alpha), \quad (1)$$

where $\mathcal{D}(\cdot)$ signifies the channel-wise splitting operator governed by the hyperparameter $\alpha \in [0, 1]$, carefully selecting the proportion of channels designated for spectral analysis. Thus, $X_l \in$

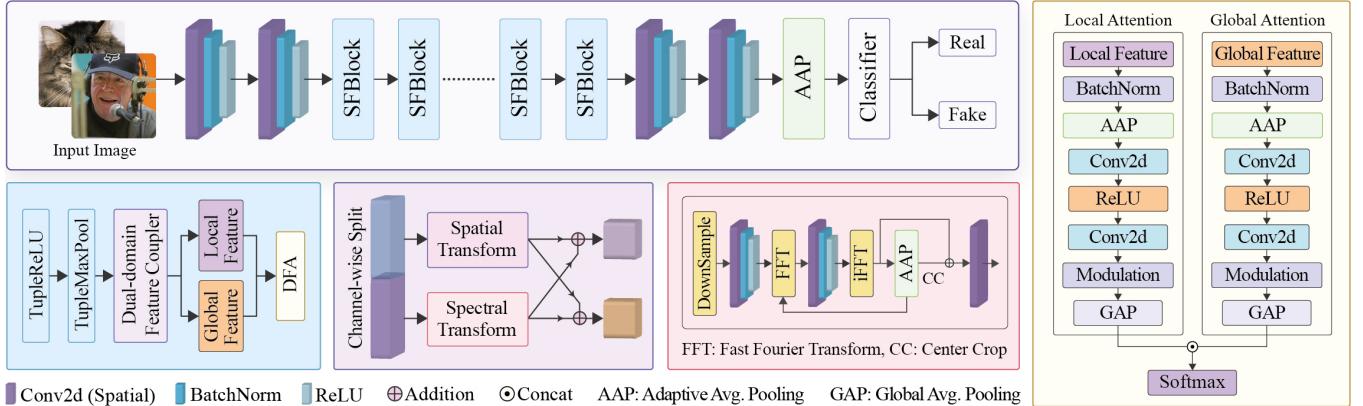


Figure 2: Overview of the proposed SpecXNet architecture for deepfake detection. The top panel presents the full pipeline, where an input image is processed through a sequence of Spectral Fusion Blocks (SFBlocks) and adaptive average pooling before final classification. The bottom-left panel illustrates the internal structure of each SFBlock, which consists of spatial convolution, max-pooling, and a Dual-Domain Feature Coupler (DDFC) module that splits the feature map into local and global branches using spatial and spectral domains, respectively. These are modulated and fused using the Dual Fourier Attention (DFA) module. The bottom-middle panel details the internal structure of the DDTC, showing how spatial and spectral pathways are handled via spatial convolution and Fourier-based spectral transforms. The bottom-right panel visualizes the internal design of this Spectral Transform, which performs downsampling, applies Fast Fourier Transform (FFT) and inverse FFT to process frequency-domain information, and incorporates a low-frequency residual stream through adaptive average pooling (AAP) and center cropping (CC) to enhance frequency-selective representation. Lastly, the right panel provides a breakdown of the DFA module, which computes cross-domain attentions from both branches, applies residual modulation, and combines them using a learned softmax-weighted fusion for final representation.

$\mathbb{R}^{(1-\alpha)C \times H \times W}$ is tailored towards capturing localized spatial intricacies, while $X_g \in \mathbb{R}^{\alpha C \times H \times W}$ is preserved explicitly for global-scale spectral domain processing.

Local Spatial Branch. The core aim of the local spatial branch is to meticulously preserve and capture the intricate spatial patterns, subtle edges, and fine textures embedded within the input data. To this end, we utilize CNN layers, renowned for their efficacy in extracting local visual patterns. Given convolutional kernels parameterized by weights W_l and biases b_l , we perform the convolution operation on the local features X_l . The precise mathematical expression is formulated as follows:

$$Z_l = W_l * X_l + b_l = \sum_i \sum_j W_l(i, j) X_l(h - i, w - j) + b_l. \quad (2)$$

Following the convolutional transformation, we implement a batch normalization (BN) procedure to stabilize the intermediate activations and accelerate convergence as follows:

$$\text{BN}(Z_l) = \gamma \frac{(Z_l - \mu_B)}{\sqrt{\sigma_B^2 + \epsilon}} + \beta, \quad (3)$$

where μ_B and σ_B respectively represent the mean and variance computed across mini-batches, γ and β denote learnable scaling and shifting parameters, and ϵ is a small numerical stability constant. Lastly, we incorporate a non-linear rectified linear unit (ReLU) activation $\sigma(\cdot)$ to introduce nonlinearity into our representation:

$$Y_l = \sigma(\text{BN}(Z_l)) = \max(0, \text{BN}(Z_l)). \quad (4)$$

The resultant feature map Y_l explicitly encodes local structures, laying a robust foundation for capturing high-resolution spatial details crucial to accurate representation learning.

Global Spectral Branch. Simultaneously, to integrate a broader receptive field and global contextual understanding, we propose an innovative spectral processing branch. This global branch explicitly leverages frequency-domain representation via the Fast Fourier Transform (FFT), fundamentally transforming spatial data into frequency-domain representations. We initiate the process by converting spatial domain features X_g into frequency domain representations X_g^F via a 2D Fourier Transform, mathematically expressed as follows:

$$X_g^F(u, v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X_g(h, w) e^{-j2\pi \left(\frac{uh}{H} + \frac{vw}{W} \right)}, \quad (5)$$

where (u, v) indicates the frequency-domain indices, and the transform effectively encodes global-level correlations across the entirety of the input spatial field into each spectral coefficient.

After obtaining spectral representations from Eq. 5, we apply a learnable modulation operation $\Phi(\cdot)$, which serves to selectively emphasize critical spectral features. This operation is implemented via an element-wise spectral filtering parameterized by learnable spectral weights W_g . Hence, the modulated spectral coefficients $X_g'F(u, v)$ are represented as follows:

$$X_g'F(u, v) = \sigma \left(\text{BN}(W_g \odot X_g^F(u, v)) \right), \quad (6)$$

where \odot denotes the Hadamard (element-wise) product, and the spectral batch normalization is employed analogously to its spatial counterpart to facilitate stable learning dynamics.

To complete the spectral processing cycle, we subsequently revert the frequency-modulated features $X'_g F(u, v)$ back into the spatial domain. This inversion is accomplished via the inverse Fourier Transform, meticulously reconstructing a spatial representation $Y_g(h, w)$ from the global-scale spectral information. Explicitly, the inverse Fourier Transform is formulated as:

$$Y_g(h, w) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} X'_g F(u, v) e^{j2\pi \left(\frac{uh}{H} + \frac{vw}{W} \right)}. \quad (7)$$

The resulting global representation Y_g thus inherently encodes comprehensive contextual information spanning the entire input, effectively complementing the local branch by capturing extensive inter-pixel dependencies that exceed the limited receptive field achievable through purely local convolutions. As illustrated in **red block** Figure 2, the spectral branch leverages both full-resolution and pooled low-frequency signals via the *Spectral Transform* module, integrating them through frequency-domain processing and subsequent refinement. Collectively, the joint processing of local and global branches delivers a synergistic integration of fine-grained local information and expansive global context, offering a robust representation for complex pattern recognition tasks.

3.2 Dual Fourier Attention (DFA)

To effectively leverage the complementary yet distinct representations learned by the local and global processing branches, we introduce a meticulously designed Dual Fourier Attention (DFA) mechanism. DFA is explicitly formulated to adaptively fuse spatially precise local features with contextually enriched global spectral representations. This fusion mechanism dynamically recalibrates feature importance across domains, significantly enhancing the discriminative capabilities of the resultant representations.

The proposed DFA commences with the compression of spatial information into compact feature descriptors via a global average pooling (GAP) operation. Given spatial feature maps $Y_l \in \mathbb{R}^{C_l \times H \times W}$ and $Y_g \in \mathbb{R}^{C_g \times H \times W}$ from the local and global branches respectively, the GAP operation aggregates spatial information into concise global descriptors $Z_l \in \mathbb{R}^{C_l}$ and $Z_g \in \mathbb{R}^{C_g}$ as follows:

$$Z_l = \mathcal{P}(Y_l) = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W Y_l(:, h, w), \quad (8)$$

$$Z_g = \mathcal{P}(Y_g) = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W Y_g(:, h, w), \quad (9)$$

where $\mathcal{P}(\cdot)$ denotes the global spatial averaging across all spatial locations for each feature channel independently. Such aggregation succinctly encapsulates dominant spatial patterns and global spectral structures, distilling the crucial representational insights from both local and global domains into lower-dimensional embeddings.

Subsequent to global pooling, these feature descriptors are then utilized to generate two attention maps, specifically tailored for bidirectional cross-domain modulation. The global attention map $A_g \in \mathbb{R}^{C_l}$ and local attention map $A_l \in \mathbb{R}^{C_g}$ are derived from the

respective feature descriptors through separate linear transformations with trainable parameters as follows:

$$A_g = \sigma(W_a Z_g + b_a), \quad (10)$$

$$A_l = \sigma(W_b Z_l + b_b), \quad (11)$$

where $W_a \in \mathbb{R}^{C_l \times C_g}$, $W_b \in \mathbb{R}^{C_g \times C_l}$, $b_a \in \mathbb{R}^{C_l}$, and $b_b \in \mathbb{R}^{C_g}$ represent learnable weights and biases respectively, initialized to optimize feature fusion during training. Here, the sigmoid activation function $\sigma(x) = 1/(1+e^{-x})$ is strategically employed to ensure that the computed attention maps A_g and A_l smoothly range between 0 and 1, providing a probabilistic interpretation of the relative importance of features from each domain.

(In practice, this attention generation is implemented via two light-weight convolutional sub-networks composed of 1×1 convolutions with intermediate non-linearity (ReLU), closely mimicking fully connected layers applied channel-wise. These convolutional pathways are applied to the GAP-reduced tensors reshaped as single-spatial pixel feature maps.)

(Furthermore, to ensure shape consistency during fusion, the attention maps A_g and A_l are upsampled via bilinear interpolation if their spatial dimensions differ from their target domains. This resizing ensures proper broadcasting during modulation.)

The cross-domain modulation phase explicitly utilizes the derived attention maps to modulate the complementary domain features through element-wise multiplications. This crucial step selectively emphasizes or suppresses features according to cross-domain feedback, effectively integrating multi-scale information into each domain's representation. The modulated feature maps are thus formulated as follows:

$$\tilde{Y}_l = Y_l \odot A_g, \quad (12)$$

$$\tilde{Y}_g = Y_g \odot A_l, \quad (13)$$

where \odot is element-wise multiplication, which dynamically adjusts the feature intensity based on domain-specific attention scores.

(Notably, in our implementation, a residual formulation is adopted where the attention output is added back to the original features, i.e., $\tilde{Y}_l = Y_l + Y_l \odot A_g$ and $\tilde{Y}_g = Y_g + Y_g \odot A_l$, to preserve identity information and stabilize gradient flow. This residual modulation enhances representation fidelity during training.)

Through this adaptive modulation, DFA robustly enables the local branch to explicitly incorporate global context and simultaneously allows the global branch to embed refined local details.

To finalize the fusion, we introduce a carefully designed adaptive weighting strategy that further optimizes the combined contribution of each modulated feature map. To achieve this, we concatenate the pooled global descriptors Z_l and Z_g into a single unified vector and subsequently compute adaptive fusion coefficients using a learnable linear mapping that is defined as follows:

$$[y_l, y_g] = \text{softmax}(W_f [Z_l; Z_g] + b_f), \quad (14)$$

where $[.; .]$ denotes concatenation, $W_f \in \mathbb{R}^{2 \times (C_l+C_g)}$ and $b_f \in \mathbb{R}^2$ are trainable parameters, and the softmax function ensures that the weights y_l and y_g represent valid probability distributions, thereby guaranteeing their sum to unity. The learned coefficients intuitively adapt to input feature characteristics, allowing the network to intelligently balance the contribution from each domain based on

the prevailing representational context. Ultimately, the final integrated representation Y_{out} is computed as a weighted combination of the modulated local and global feature maps using these adaptive fusion weights:

$$Y_{\text{out}} = \gamma_l \tilde{Y}_l + \gamma_g \tilde{Y}_g. \quad (15)$$

As illustrated by the orange block in Figure 2, the DFA mechanism captures the interactions between local and global features through symmetric attention pathways, followed by residual modulation and adaptive fusion driven by global descriptors. This meticulous formulation endows DFA with the capacity to dynamically recalibrate and optimally merge local and global representations in a context-sensitive manner. Consequently, the DFA enhances the representational capability of SpecXNet, facilitating superior generalization performance and enriched feature expressiveness.

3.3 SpecXNet Backbone

XceptionNet Integration Procedure. We integrate our proposed DDFC and DFA mechanisms into a refined XceptionNet architecture. The primary innovation behind XceptionNet lies in the systematic employment of depthwise separable convolutions, which factorize standard convolutional operations into spatially separate depthwise and pointwise convolutions that can be delineated as follows:

$$Y_{\text{sep}} = W_p * (W_d \otimes X), \quad (16)$$

where X denotes the input feature map, W_d represents a depthwise convolution kernel applied independently to each channel, \otimes denotes depthwise convolution, and W_p indicates a pointwise convolution kernel responsible for inter-channel interactions.

Building upon this efficient convolutional design, we carefully embed our dual-domain modules: local spatial branch and global spectral branch, interconnected by our proposed DFA, into XceptionNet blocks. Each enhanced Xception block systematically processes the input feature maps $X^i \in \mathbb{R}^{C \times H \times W}$ through a balanced interplay between local and global transformations. Specifically, for the i -th modified block, input features are initially decomposed into local (X_l^i) and global (X_g^i) components using the same channel-wise decomposition described in Eq. 1.

The local features X_l^i are processed using depthwise separable convolutions, batch normalization, and nonlinear activation, following a structure similar to Eqs. 2–4.

Simultaneously, the global spectral features X_g^i undergo frequency domain transformation, modulation, and reconstruction. The forward Fourier Transform is computed as shown in Eq. 5, followed by learnable spectral modulation via Eq. 6, and finally, the inverse Fourier Transform is applied as per Eq. 7 to retrieve spatial-domain global features Y_g^i .

The two representations, Y_l^i from the spatial branch and Y_g^i from the spectral branch, are subsequently fused via the Dual Fourier Attention (DFA) mechanism. This mechanism adaptively modulates cross-domain features and integrates them using learned fusion weights, as formulated in Eqs. 10–15.

This harmonious integration within the modified XceptionNet architecture enables the network to simultaneously benefit from high-resolution spatial cues and holistic spectral context, offering both computational efficiency and powerful generalization capability across diverse deepfake scenarios.

Optimization Procedure. Optimization of our integrated dual-domain XceptionNet architecture is performed rigorously using standard stochastic gradient descent (SGD), equipped with momentum and adaptive learning rate scheduling. To formally define our learning process, consider a labeled training set $\{(X_j, Y_j)\}_{j=1}^N$ consisting of N pairs, where X_j denotes an input image and Y_j corresponds to its associated ground-truth label. The network's parameters θ , comprising convolutional kernels, attention weights, and modulation parameters are iteratively refined to minimize an empirical risk function \mathcal{L} as follows:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{j=1}^N \ell(Y_{\text{pred}}(X_j; \theta), Y_j) + \lambda R(\theta), \quad (17)$$

where ℓ typically denotes cross-entropy loss for classification tasks, while $R(\theta)$ represents a regularization term penalizing overly complex parameterizations and enforcing stability during training, with weight λ balancing the empirical loss and model complexity.

At each optimization iteration t , the network parameters are updated via gradient descent steps, computed as follows:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_{\theta} \mathcal{L}(\theta_t), \quad (18)$$

where η_t is the learning rate dynamically adjusted using cosine annealing or step-wise decay to ensure convergence and mitigate local minima entrapment. Moreover, we incorporate momentum-based updates to accelerate convergence as follows:

$$v_{t+1} = \mu v_t - \eta_t \nabla_{\theta} \mathcal{L}(\theta_t), \quad \theta_{t+1} = \theta_t + v_{t+1}, \quad (19)$$

where a momentum factor $\mu \in [0, 1]$ controlling inertia from previous gradients.

Through meticulous backpropagation, gradients are carefully computed, taking into account intricate interactions between local convolutional layers, global spectral modules, and the DFA attention mechanism. This comprehensive optimization procedure encourages the XceptionNet architecture to learn discriminative, multi-domain representations, ultimately facilitating significant performance improvements in challenging vision tasks.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Implementation Details

Datasets. To evaluate the robustness and generalizability of our proposed framework, we conduct experiments on a diverse collection of real and synthetic datasets. For real-world relevance, we adopt widely-used datasets introduced by Wu et al. [66] and further augment our benchmarks with self-curated data synthesized via SOTA generative models. Specifically, fake samples are generated using Stable Diffusion v1.4 [47, 56], DreamBooth [44], and Latent Diffusion models from CompVis, thereby ensuring coverage of contemporary diffusion-based generation techniques.

Our evaluation encompasses synthetic data produced by eleven distinct generation pipelines, including ProGAN [32], StyleGAN2 [34], StyleGAN3 [74], BigGAN [6], EG3D [8], Taming Transformers [21], DALL-E 2 [55], GLIDE [51], Latent Diffusion [25], Guided Diffusion [20], and Stable Diffusion v1.4. Corresponding real image distributions are sourced from large-scale and diverse datasets, namely ImageNet [19], COCO [41], and Danbooru [1].

To further assess the model’s effectiveness under real-world conditions, we introduce a new practical benchmark denoted as T_{Gen} , comprising 1,000 images per category from DreamBooth, Midjourney v4 and v5 [26], NightCafe [58], StableAI, and Yijian [43]. Prompt engineering and generation details for these samples are presented in the Appendix. Additionally, to establish a comprehensive generalization analysis, we evaluate our model on the GenImage benchmark [73], a dataset designed to test deepfake detectors under varied synthetic domains. For evaluating diffusion-based generation, we also sample 1,000 images each from SDXL [53] and DiffusionDB [62] to ensure a robust and diverse test suite covering different model classes. Our experimental suite includes both object-centric and face-centric manipulations, spanning indoor, outdoor, and natural scenes to ensure domain diversity. The complete training corpus contains 470K real and 410K synthetic images. During testing, we uniformly sample 5,000 images per synthetic category, yielding a total of 62K fake and 62K real images in the test set. Detailed dataset splits are reported in Table 1.

Experimental Setup. All experiments are conducted using PyTorch (v3.6) on a system equipped with four NVIDIA Titan RTX GPUs, accelerated via CUDA 11.3. Training is performed with a batch size of 48, and all images are resized to a fixed resolution of 224×224 pixels. We use the standard cross-entropy loss to optimize classification, coupled with the Adam optimizer initialized at a learning rate of 0.1. The training process incorporates a composite learning rate schedule: an initial 5-epoch warm-up phase is followed by a cosine annealing strategy to gradually decay the learning rate. Additionally, learning rate reductions by a factor of 10 are applied at epochs 30, 60, and 80 to promote convergence. Model performance is assessed using three evaluation metrics: accuracy, Area Under the ROC Curve (AUC), and mean Average Precision (mAP) to offer a well-rounded view of detection capability.

4.2 Comparison with SOTA Methods

We extensively evaluated our approach on different datasets using our method which is enhanced with dual-domain processing, rather than relying on pre-trained models such as CLIP. Unlike prior methods that heavily depend on off-the-shelf feature extractors, our approach integrates spectral-spatial learning via a customized spectral cross-attentional framework. As shown in the last row in Table 2, our method achieves the highest average AUC and accuracy compared to existing SOTA approaches: Wang et al. [64], Chandrasegaran et al. [9], Chai et al. [7], Grag et al. [22], Xu Zhang et al. [70], and UGAD [3]. We find that our approach outperforms these methods in the all of cases.

In addition, recent works have explored generalization in fake image detection using vision-language models. Davide et al. [18] investigate the use of CLIP features for detecting AI-generated content and demonstrate strong generalization across models like DALL-E 3 and MidjourneyV5 with minimal training data. Similarly, Ojha et al. [50] propose a universal fake image detector that generalizes across multiple generative models, while “CLIPping the Deception” [37] adapts vision-language models for universal deepfake detection, showing their applicability across image domains. Corvi et al. [16] focus specifically on diffusion models, evaluating detection methods tailored to this emerging class of generators.

Table 1: Summary of datasets used for training, evaluation, and generalization testing. The “Train” and “Test” columns indicate the dataset’s usage in model training and evaluation, respectively. The number of test images is set to 5,000 fewer than the training samples for entries with \checkmark in both columns. All real image families include 20,000 samples in the test set. The T_{Gen} column indicates inclusion in the practical generalization benchmark.

Family	Method	#Images	Train	Test	T_{Gen}
Real	ImageNet	140k	\checkmark	\checkmark	\times
Real	MS COCO	120k	\checkmark	\checkmark	\times
Real	LSUN	120k	\checkmark	\times	\times
Real	Danbooru & Artist	110k	\checkmark	\checkmark	\times
GAN	ProGAN	210k	\checkmark	\checkmark	\times
GAN	StyleGAN2	5k	\times	\checkmark	\times
GAN	StyleGAN3	5k	\times	\checkmark	\times
GAN	BigGAN	5k	\times	\checkmark	\times
GAN	Eg3D	5k	\times	\checkmark	\times
Transformer	Taming Transformer	5k	\times	\checkmark	\times
Diffusion	GLIDE	6k	\times	\checkmark	\times
Diffusion	Stable Diffusion V1.4	210k	\checkmark	\checkmark	\times
Diffusion	Latent Diffusion	7k	\times	\checkmark	\times
Diffusion	DALL-E 2	7k	\times	\checkmark	\times
Diffusion	Guided Diffusion	5k	\times	\checkmark	\times
Diffusion	SDXL	1k	\times	\checkmark	\times
Diffusion	DiffusionDB	1k	\times	\checkmark	\times
Diffusion	DreamBooth	1k	\times	\checkmark	\checkmark
Diffusion	MidjourneyV4	1k	\times	\checkmark	\checkmark
Diffusion	MidjourneyV5	1k	\times	\checkmark	\checkmark
Diffusion	NightCafe	1k	\times	\checkmark	\checkmark
Diffusion	StableAI	1k	\times	\checkmark	\checkmark
Diffusion	Yijian	1k	\times	\checkmark	\checkmark

These studies collectively emphasize that large domain-specific datasets are not essential, and lightweight detection pipelines can yield competitive results. We include a comparative evaluation against these approaches in Table 3, where our method consistently demonstrates superior performance in accuracy for distinguishing real and fake content except Corvi et al. in LDM method. For this evaluation, we use the GenImage dataset [73], which contains a total of 24,000 synthetic images evenly split across eight generative families. For deepfake benchmarks, we adopt the FaceForensics++ (FF++) dataset [57], utilizing 100 test videos each for Deepfakes and FaceSwap, corresponding to approximately 10,000 frames per manipulation type. Additionally, we sample 1,000 images each from diffusion-based models including Guided Diffusion, SDXL, and DiffusionDB to evaluate generalization performance across diverse diffusion datasets.

4.3 Ablation Studies

Performance Scaling with Different Architecture. Table 4 reports the incremental gains from integrating the proposed modules across ResNet [24] and XceptionNet backbones. ResNet50 sees an increase from 69.0% to 84.1%, while ResNet101 and ResNet152 improve from 78.2% and 79.9% to 90.2% and 93.0%, respectively. XceptionNet shows the strongest performance, reaching 96.4%. These results indicate that deeper networks benefit more from spectral-spatial modeling and attention-based fusion. The consistent improvements across architectures confirm that the proposed modules generalize well, significantly enhancing representational power.

Table 2: Benchmark comparison of deepfake detection methods evaluated across a comprehensive suite of GAN-based and diffusion-based generative models, including datasets derived from ImageNet, COCO, artist-rendered sources, and Danbooru. The table contrasts the performance of several SOTA baselines with our proposed SpecXNet, all trained using ProGAN and Stable Diffusion V1.4 for consistent evaluation. Here, **Bold indicates best performance, while underlined denotes second-best.**

Datasets		Grag			CR			Wang			Zhang			PatchFor			UGAD			SpecXNet (Ours)		
Real	Fake	AUC	mAP	Acc	AUC	mAP	Acc															
ImageNet + COCO	Real	.745	<u>.826</u>	<u>.796</u>	<u>.725</u>	<u>.657</u>	<u>.534</u>	<u>.858</u>	<u>.843</u>	<u>.795</u>	<u>.485</u>	<u>.621</u>	<u>.497</u>	<u>.653</u>	<u>.581</u>	<u>.504</u>	<u>.951</u>	<u>.952</u>	<u>.936</u>	.981	.965	.972
	BigGAN	<u>.745</u>	<u>.826</u>	<u>.796</u>	<u>.725</u>	<u>.657</u>	<u>.534</u>	<u>.858</u>	<u>.843</u>	<u>.795</u>	<u>.485</u>	<u>.621</u>	<u>.497</u>	<u>.653</u>	<u>.581</u>	<u>.504</u>	<u>.951</u>	<u>.952</u>	<u>.936</u>	.981	.965	.972
	StyleGAN2	<u>.858</u>	<u>.950</u>	<u>.912</u>	<u>.870</u>	<u>.590</u>	<u>.558</u>	<u>.899</u>	<u>.765</u>	<u>.728</u>	<u>.505</u>	<u>.540</u>	<u>.518</u>	<u>.736</u>	<u>.770</u>	<u>.508</u>	<u>.967</u>	<u>.957</u>	<u>.928</u>	<u>.975</u>	<u>.966</u>	<u>.945</u>
	StyleGAN3	<u>.908</u>	<u>.895</u>	<u>.854</u>	<u>.869</u>	<u>.645</u>	<u>.615</u>	<u>.901</u>	<u>.785</u>	<u>.754</u>	<u>.519</u>	<u>.550</u>	<u>.529</u>	<u>.767</u>	<u>.790</u>	<u>.502</u>	<u>.921</u>	<u>.915</u>	<u>.877</u>	<u>.942</u>	<u>.930</u>	<u>.902</u>
	ProGAN	<u>.833</u>	<u>.810</u>	<u>.772</u>	<u>.794</u>	<u>.685</u>	<u>.654</u>	<u>.880</u>	<u>.865</u>	<u>.815</u>	<u>.485</u>	<u>.510</u>	<u>.497</u>	<u>.653</u>	<u>.680</u>	<u>.504</u>	<u>.987</u>	<u>.980</u>	<u>.957</u>	<u>.993</u>	<u>.986</u>	<u>.962</u>
	EG3D	<u>.793</u>	<u>.710</u>	<u>.668</u>	<u>.856</u>	<u>.575</u>	<u>.537</u>	<u>.860</u>	<u>.840</u>	<u>.799</u>	<u>.606</u>	<u>.635</u>	<u>.589</u>	<u>.819</u>	<u>.850</u>	<u>.498</u>	<u>.872</u>	<u>.860</u>	<u>.834</u>	<u>.974</u>	<u>.965</u>	<u>.942</u>
	DALL-E 2	<u>.516</u>	<u>.580</u>	<u>.552</u>	<u>.522</u>	<u>.555</u>	<u>.520</u>	<u>.586</u>	<u>.610</u>	<u>.560</u>	<u>.650</u>	<u>.675</u>	<u>.620</u>	<u>.584</u>	<u>.612</u>	<u>.497</u>	<u>.941</u>	<u>.965</u>	<u>.872</u>	<u>.962</u>	<u>.973</u>	<u>.945</u>
	GLIDE	<u>.574</u>	<u>.620</u>	<u>.588</u>	<u>.624</u>	<u>.655</u>	<u>.528</u>	<u>.608</u>	<u>.640</u>	<u>.600</u>	<u>.525</u>	<u>.550</u>	<u>.531</u>	<u>.715</u>	<u>.745</u>	<u>.510</u>	<u>.936</u>	<u>.965</u>	<u>.927</u>	<u>.950</u>	<u>.970</u>	<u>.941</u>
	Latent Diffusion	<u>.863</u>	<u>.705</u>	<u>.675</u>	<u>.844</u>	<u>.880</u>	<u>.907</u>	<u>.749</u>	<u>.784</u>	<u>.650</u>	<u>.463</u>	<u>.485</u>	<u>.479</u>	<u>.652</u>	<u>.682</u>	<u>.506</u>	<u>.970</u>	<u>.951</u>	<u>.921</u>	<u>.975</u>	<u>.980</u>	<u>.943</u>
	Taming Transformer	<u>.710</u>	<u>.722</u>	<u>.692</u>	<u>.757</u>	<u>.787</u>	<u>.703</u>	<u>.943</u>	<u>.973</u>	<u>.652</u>	<u>.791</u>	<u>.821</u>	<u>.790</u>	<u>.741</u>	<u>.771</u>	<u>.610</u>	<u>.950</u>	<u>.980</u>	<u>.876</u>	<u>.978</u>	<u>.985</u>	<u>.940</u>
Artist + Danbooru	Stable DiffusionV1.4	<u>.580</u>	<u>.615</u>	<u>.592</u>	<u>.578</u>	<u>.601</u>	<u>.523</u>	<u>.577</u>	<u>.598</u>	<u>.609</u>	<u>.415</u>	<u>.476</u>	<u>.444</u>	<u>.731</u>	<u>.789</u>	<u>.698</u>	<u>.922</u>	<u>.952</u>	<u>.937</u>	<u>.955</u>	<u>.970</u>	<u>.945</u>
	Guided Diffusion	<u>.588</u>	<u>.607</u>	<u>.577</u>	<u>.584</u>	<u>.614</u>	<u>.520</u>	<u>.566</u>	<u>.596</u>	<u>.652</u>	<u>.491</u>	<u>.521</u>	<u>.491</u>	<u>.691</u>	<u>.721</u>	<u>.510</u>	<u>.925</u>	<u>.945</u>	<u>.915</u>	<u>.958</u>	<u>.965</u>	<u>.948</u>
	BigGAN	<u>.949</u>	<u>.913</u>	<u>.883</u>	<u>.892</u>	<u>.922</u>	<u>.958</u>	<u>.946</u>	<u>.976</u>	<u>.984</u>	<u>.705</u>	<u>.735</u>	<u>.733</u>	<u>.857</u>	<u>.887</u>	<u>.725</u>	<u>.942</u>	<u>.972</u>	<u>.956</u>	<u>.986</u>	<u>.984</u>	<u>.978</u>
	StyleGAN2	<u>.951</u>	<u>.913</u>	<u>.883</u>	<u>.911</u>	<u>.941</u>	<u>.899</u>	<u>.969</u>	<u>.999</u>	<u>.972</u>	<u>.770</u>	<u>.800</u>	<u>.713</u>	<u>.883</u>	<u>.913</u>	<u>.702</u>	<u>.985</u>	<u>.965</u>	<u>.982</u>	<u>.989</u>	<u>.986</u>	<u>.985</u>
	GANStyleGAN3	<u>.978</u>	<u>.981</u>	<u>.951</u>	<u>.966</u>	<u>.996</u>	<u>.967</u>	<u>.969</u>	<u>.999</u>	<u>.972</u>	<u>.440</u>	<u>.470</u>	<u>.338</u>	<u>.718</u>	<u>.748</u>	<u>.500</u>	<u>.983</u>	<u>.955</u>	<u>.970</u>	<u>.990</u>	<u>.987</u>	<u>.986</u>
	ProGAN	<u>.970</u>	<u>.974</u>	<u>.899</u>	<u>.935</u>	<u>.980</u>	<u>.986</u>	<u>.952</u>	<u>.980</u>	<u>.991</u>	<u>.992</u>	<u>.997</u>	<u>.835</u>	<u>.976</u>	<u>.988</u>	<u>.649</u>	<u>.992</u>	<u>.997</u>	<u>.992</u>	<u>.995</u>	<u>.998</u>	<u>.994</u>
	EG3D	<u>.823</u>	<u>.847</u>	<u>.722</u>	<u>.826</u>	<u>.850</u>	<u>.787</u>	<u>.905</u>	<u>.925</u>	<u>.943</u>	<u>.803</u>	<u>.828</u>	<u>.896</u>	<u>.500</u>	<u>.525</u>	<u>.784</u>	<u>.962</u>	<u>.984</u>	<u>.985</u>	<u>.986</u>	<u>.989</u>	<u>.987</u>
	DALL-E 2	<u>.778</u>	<u>.803</u>	<u>.616</u>	<u>.782</u>	<u>.808</u>	<u>.827</u>	<u>.746</u>	<u>.770</u>	<u>.737</u>	<u>.933</u>	<u>.960</u>	<u>.821</u>	<u>.525</u>	<u>.550</u>	<u>.500</u>	<u>.955</u>	<u>.980</u>	<u>.898</u>	<u>.963</u>	<u>.985</u>	<u>.961</u>
	GLIDE	<u>.827</u>	<u>.850</u>	<u>.668</u>	<u>.754</u>	<u>.790</u>	<u>.845</u>	<u>.754</u>	<u>.780</u>	<u>.814</u>	<u>.819</u>	<u>.845</u>	<u>.734</u>	<u>.966</u>	<u>.990</u>	<u>.503</u>	<u>.974</u>	<u>.970</u>	<u>.957</u>	<u>.981</u>	<u>.986</u>	<u>.963</u>
	Stable DiffusionV1.4	<u>.765</u>	<u>.800</u>	<u>.659</u>	<u>.767</u>	<u>.807</u>	<u>.726</u>	<u>.901</u>	<u>.931</u>	<u>.882</u>	<u>.742</u>	<u>.885</u>	<u>.701</u>	<u>.832</u>	<u>.865</u>	<u>.659</u>	<u>.962</u>	<u>.965</u>	<u>.936</u>	<u>.984</u>	<u>.988</u>	<u>.967</u>
	Guided Diffusion	<u>.869</u>	<u>.885</u>	<u>.670</u>	<u>.818</u>	<u>.850</u>	<u>.897</u>	<u>.716</u>	<u>.735</u>	<u>.681</u>	<u>.793</u>	<u>.810</u>	<u>.769</u>	<u>.831</u>	<u>.850</u>	<u>.644</u>	<u>.988</u>	<u>.995</u>	<u>.974</u>	<u>.993</u>	<u>.997</u>	<u>.982</u>
	Latent Diffusion	<u>.863</u>	<u>.695</u>	<u>.675</u>	<u>.844</u>	<u>.875</u>	<u>.908</u>	<u>.749</u>	<u>.770</u>	<u>.650</u>	<u>.785</u>	<u>.805</u>	<u>.721</u>	<u>.843</u>	<u>.865</u>	<u>.678</u>	<u>.971</u>	<u>.980</u>	<u>.959</u>	<u>.985</u>	<u>.992</u>	<u>.975</u>
	Taming Transformer	<u>.865</u>	<u>.670</u>	<u>.651</u>	<u>.878</u>	<u>.900</u>	<u>.931</u>	<u>.966</u>	<u>.985</u>	<u>.915</u>	<u>.866</u>	<u>.890</u>	<u>.796</u>	<u>.694</u>	<u>.715</u>	<u>.545</u>	<u>.969</u>	<u>.945</u>	<u>.960</u>	<u>.987</u>	<u>.993</u>	<u>.982</u>
Average		<u>.799</u>	<u>.770</u>	<u>.729</u>	<u>.787</u>	<u>.815</u>	<u>.746</u>	<u>.815</u>	<u>.850</u>	<u>.781</u>	<u>.662</u>	<u>.690</u>	<u>.637</u>	<u>.738</u>	<u>.770</u>	<u>.570</u>	<u>.957</u>	<u>.963</u>	<u>.935</u>	<u>.985</u>	<u>.989</u>	<u>.978</u>

Table 3: Comprehensive comparison of deepfake detection performance across diverse generative models. Evaluation is conducted on GenImage (covering both GAN-based and diffusion-based generators), recent diffusion techniques, and traditional face manipulation datasets (FF++). The table contrasts the performance of several SOTA baselines with our proposed SpecXNet. All methods, including our SpecXNet, are trained using ProGAN and Stable Diffusion V1.4 to ensure fairness in evaluation. Accuracy is reported across each category, with the final column representing the overall average across all benchmarks.

Method	GenImage								Diffusion Methods						FF++		Average
	BigGAN	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	Midjourney	Glide	Guided	LD	SDXL	DiffusionDB	DeepFakes	FaceSwap		
Davide et al.	<u>.868</u>	<u>.851</u>	<u>.806</u>	<u>.824</u>	<u>.843</u>	<u>.837</u>	<u>.819</u>	<u>.826</u>	<u>.564</u>	<u>.551</u>	<u>.611</u>	<u>.629</u>	<u>.580</u>	<u>.527</u>	<u>.497</u>	<u>.706</u>	
Corvi et al.	<u>.883</u>	<u>.897</u>	<u>.874</u>	<u>.863</u>	<u>.859</u>	<u>.832</u>	<u>.821</u>	<u>.847</u>	<u>.598</u>	<u>.509</u>	<u>.973</u>	<u>.891</u>	<u>.901</u>	<u>.593</u>	<u>.501</u>	<u>.786</u>	
Ojha et al.	<u>.902</u>	<u>.910</u>	<u>.896</u>	<u>.851</u>	<u>.889</u>	<u>.871</u>	<u>.865</u>	<u>.883</u>	<u>.754</u>	<u>.895</u>	<u>.892</u>	<u>.822</u>	<u>.848</u>	<u>.799</u>	<u>.602</u>	<u>.825</u>	
Khan et al.	<u>.928</u>	<u>.930</u>	<u>.917</u>	<u>.905</u>	<u>.899</u>	<u>.882</u>	<u>.898</u>	<u>.906</u>	<u>.915</u>	<u>.929</u>	<u>.847</u>	<u>.778</u>	<u>.751</u>	<u>.784</u>	<u>.747</u>	<u>.868</u>	
Ours	<u>.965</u>	<u>.978</u>	<u>.963</u>	<u>.932</u>	<u>.958</u>	<u>.949</u>	<u>.946</u>	<u>.958</u>	<u>.960</u>	<u>.937</u>	<u>.947</u>	<u>.902</u>	<u>.904</u>	<u>.832</u>	<u>.824</u>	<u>.930</u>	

Effectiveness of Our Proposed Components. The proposed DDFC contributes significantly to the model’s ability to discern subtle yet global inconsistencies present in manipulated imagery. Unlike conventional convolutional backbones that predominantly capture spatial features, DDFC explicitly decomposes the input into local and spectral pathways, enabling the capture of periodic artifacts and frequency distortions common in synthetic content. Empirical results in Table 4 reveal that incorporating DDFC alone leads to consistent improvements across all backbone networks, with average gains exceeding 10% in many cases. These findings indicate that spectral cues when modeled through structured decomposition, provide a powerful inductive bias for forgery detection.

The DFA mechanism serves as a critical bridge between the spatial and spectral domains by dynamically modulating their interactions. Rather than naively concatenating features from distinct domains, DFA introduces a cross-attention mechanism that selectively amplifies or suppresses local and global features based on

content-aware importance. This results in more discriminative representations and enhanced robustness to diverse manipulations. When applied independently, DFA builds upon the DDFC’s output by refining the fusion process, yielding measurable performance boosts, especially in deeper architectures where semantic granularity is richer.

When jointly applied, DDFC and DFA form a synergistic framework that capitalizes on their respective strengths: DDFC’s ability to encode orthogonal domain features and DFA’s capacity for adaptive feature alignment. The dual-branch formulation of DDFC ensures comprehensive representation coverage, while DFA optimally blends these features into a unified embedding space. As reflected in the substantial accuracy gains across all evaluated architectures including a peak performance of 96.4% on XceptionNet, this combination yields a model that is not only highly performant but also generalizes well to unseen manipulations. The architecture

Table 4: Impact of ResNet architecture and SpecXNet components (DDFC and DFA) across different generation methods. SG2 denotes StyleGAN2, LD refers to Latent Diffusion. ‘Average’ indicates mean performance across all test datasets.

Backbone	Configuration	Accuracy on Generation Methods				Average
		ImageNet	ImageNet	Artist+Danbooru	Artist+Danbooru	
		+ SG2	+ DALL-E 2	+ GLIDE	+ LD	
ResNet50	None	.713	.679	.697	.681	.690
ResNet50	DDFC	.837	.792	.814	.814	.817
ResNet50	DDFC + DFA	.851	.807	.836	.854	.841
ResNet101	None	.796	.751	.792	.764	.782
ResNet101	DDFC	.895	.842	.901	.909	.887
ResNet101	DDFC + DFA	.901	.848	.929	.912	.902
ResNet152	None	.832	.778	.798	.788	.799
ResNet152	DDFC	.905	.857	.912	.918	.928
XceptionNet	None	.847	.794	.816	.802	.813
XceptionNet	DDFC	.912	.888	.921	.927	.936
XceptionNet	DDFC + DFA	.945	.945	.963	.975	.964

Table 5: Evaluation of cross-domain generalization on the practical benchmark dataset T_{Gen} , measured by classification accuracy. Each model is trained using our dataset and evaluated on six distinct generative subsets. Wang-0.1 and Wang-0.5 denote models trained using 10% and 50% data augmentation, respectively.

Method	$T_{\text{gen}}\text{-Accuracy}$						Average
	Dream Booth	Midjourney V4	Midjourney V5	Night Cafe	Stable AI	Yi Jian	
Wang-0.5	.839	.829	.795	.811	.800	.729	.801
Wang-0.1	.847	.861	.881	.867	.832	.715	.833
CR	.764	.751	.725	.790	.770	.582	.731
Grag	.655	.682	.665	.731	.729	.537	.667
Ours	.966	.929	.893	.952	.948	.821	.900

demonstrates the efficacy of dual-domain modeling paired with dynamic attention in advancing the state of deepfake detection. **Generalization Capability on Practical Dataset T_{Gen} .** To assess the robustness of our model under real-world distribution shifts, we evaluate generalization performance on the T_{Gen} dataset, which comprises diverse samples generated from DreamBooth, Midjourney V4/V5, NightCafe, StableAI, and YiJian. As shown in Table 5, our model consistently outperforms prior SOTA methods across all six subdomains, achieving an average accuracy of 90.0%. Compared to the strongest baseline (Wang-0.5), which records 80.1%, our approach yields a substantial gain of 9.9%. Notably, our model surpasses others even on challenging generators such as YiJian and Midjourney V5, reflecting strong resilience to stylistic and semantic variation. This superior performance highlights the effectiveness of our dual-domain representation and adaptive attention mechanism in capturing generalizable forgery patterns.

Robustness Against Post-Processing Artifacts. To evaluate the robustness of our detector under real-world post-processing operations, we measure its generalization ability across various common augmentations on the practical dataset T_{Gen} . As visualized in Figure 3, we analyze eleven augmentation strategies including *Blur*, *JPEG compression*, their combinations *Blur+JPEG (0.5)*

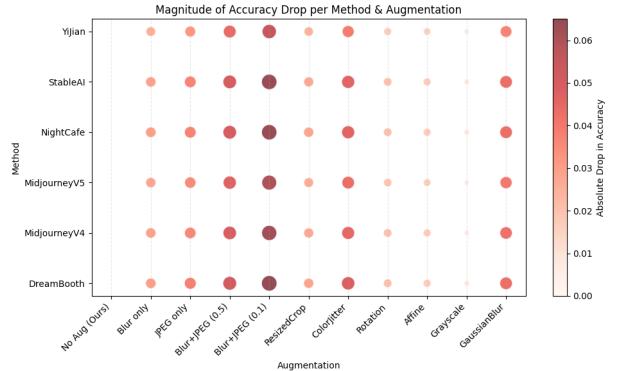


Figure 3: Effect of augmentations on detector performance across the T_{Gen} dataset. Models trained on ProGAN are evaluated on unseen generators. While most augmentations enhance robustness, certain models like MidjourneyV5 show reduced accuracy under specific perturbations.

and *Blur+JPEG (0.1)*, as well as *ResizedCrop*, *ColorJitter*, *Rotation*, *Affine*, *Grayscale*, and *GaussianBlur*, each compared against the *No Augmentation (Ours)* baseline. Our model maintains high accuracy across all cases, but the most substantial drops are observed under dual-frequency perturbations such as *Blur+JPEG*, where DreamBooth falls from 0.966 to 0.915 and 0.901 under 0.5 and 0.1 probabilities, respectively. Similar trends are seen in MidjourneyV4 and MidjourneyV5, both experiencing over 6% accuracy loss, confirming their sensitivity to low-frequency corruptions. In contrast, geometric and photometric transformations, including *Rotation* and *ColorJitter*, induce relatively minor performance degradation (typically under 2–3%). YiJian consistently remains the most challenging generator, with a baseline of 0.821 and a minimum of 0.766 under *Blur+JPEG (0.1)*, suggesting that the generative artifacts are more subtle and harder to detect. Despite these variations, our dual-domain model exhibits remarkable resilience, retaining over 90% accuracy in most configurations. This robustness can be attributed to the architecture and Dual Fourier Attention (DFA), which together capture compression-induced spectral shifts and spatial inconsistencies. The spectral branch enhances robustness against frequency-domain distortions, while the spatial branch absorbs geometric perturbations, and the content-aware DFA dynamically fuses both. These results emphasize that our architecture not only detects forgeries across different generators but also generalizes effectively under real-world post-processing, making it suitable for deployment in unconstrained environments.

Overall, the synergy of DDFC and DFA modules enhances architectural scalability and drives generalization across varied deepfake generation methods with different augmentation techniques.

5 CONCLUSION

In this work, we introduced the SpecXNet, a novel dual-domain deepfake detection framework that integrates spatial and spectral cues through a unified convolutional architecture. By explicitly decomposing feature representations into local spatial and global spectral branches, and fusing them via the proposed DFA mechanism,

DDFC effectively captures both fine-grained textural anomalies and long-range frequency inconsistencies inherent in synthetic media. Our extensive evaluations across diverse benchmarks, including GAN, diffusion, and real-world generative models, demonstrate that our proposed SpecXNet achieves SOTA performance in accuracy and generalization, even under challenging post-processing perturbations. These results highlight the critical importance of joint spatial-spectral modeling and content-aware attention for robust and scalable deepfake detection.

REFERENCES

- [1] Danbooru AI. 2021. *DanBooru Dataset*. <https://gwmn.net/danbooru2021> [Accessed on 12-04-2025].
- [2] Inzamamul Alam, Md Tanvir Islam, and Simon S Woo. 2025. Saliency-Aware Diffusion Reconstruction for Effective Invisible Watermark Removal. In *Companion Proceedings of the ACM on Web Conference 2025*. 849–853.
- [3] Inzamamul Alam, Muhammad Shahid Muneer, and Simon S Woo. 2024. UGAD: Universal Generative AI Detector utilizing Frequency Fingerprints. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4332–4340.
- [4] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. 2024. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 719–728.
- [5] Young Oh Bang and Simon S Woo. 2021. DA-FDFtNet: dual attention fake detection fine-tuning network to detect various AI-generated fake images. *arXiv preprint arXiv:2112.12001* (2021).
- [6] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).
- [7] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. 2020. What makes fake images detectable? understanding properties that generalize. In *Computer vision-ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVI 16*. Springer, 103–120.
- [8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16123–16133.
- [9] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Alexander Binder, and Ngaiman Cheung. 2022. Discovering transferable forensic features for cnn-generated images detection. In *European Conference on Computer Vision*. Springer, 671–689.
- [10] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. 2022. 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*. 18710–18719.
- [11] Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, Rongrong Ji, et al. 2024. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems* 37 (2024), 101474–101497.
- [12] Beomsang Cho, Binh M Le, Jiwon Kim, Simon Woo, Shahroz Tariq, Alsharif Abuadba, and Kristen Moore. 2023. Towards understanding of deepfake videos in the wild. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4530–4537.
- [13] Andrea Ciamarra, Roberto Caldelli, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2024. Deepfake detection by exploiting surface anomalies: the SurFake approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1024–1033.
- [14] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2023. Fakewatcher: detection of synthetic portrait videos using biological signals. US Patent 11,687,778.
- [15] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 973–982.
- [16] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [17] Davide Cozzolino, Diego Gragnaniello, Giovanni Poggi, and Luisa Verdoliva. 2021. Towards universal GAN image detection. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 1–5.
- [18] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. 2024. Raising the Bar of AI-generated Image Detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [20] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- [21] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.
- [22] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. 2021. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*. IEEE, 1–6.
- [23] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2024. Mastering deepfake detection: A cutting-edge approach to distinguish GAN and diffusion-model images. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 11 (2024), 1–24.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [26] David Holz. 2022, July 12. *Midjourney: Expanding the Imaginative Powers*. <https://www.midjourney.com/home> [Accessed on 12-04-2025].
- [27] Seunghoo Hong, Juhun Lee, and Simon S Woo. 2024. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 21143–21151.
- [28] Md Tanvir Islam, Ik Hyun Lee, Ahmed Ibrahim Alzahrani, and Khan Muhammad. 2025. MEXFIC: A meta ensemble eXplainable approach for AI-synthesized fake image classification. *Alexandria Engineering Journal* 116 (2025), 351–363.
- [29] Hyeonseong Jeon, Youngoh Bang, and Simon S Woo. 2019. Faketalkerdetect: Effective and practical realistic neural talking head detection with a highly unbalanced dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.
- [30] Hyeonseong Jeon, Youngoh Bang, and Simon S Woo. 2020. Fdfnet: Facing off fake images using fake detection fine-tuning network. In *IFIP international conference on ICT systems security and privacy protection*. Springer, 416–430.
- [31] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. 2022. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 1060–1068.
- [32] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).
- [33] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems* 33 (2020), 12104–12114.
- [34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.
- [35] Hasam Khalid and Simon S Woo. 2020. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 656–657.
- [36] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. 2023. Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms. *IEEE Access* 12 (2023), 1880–1908.
- [37] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. 2024. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*. 1006–1015.
- [38] Konwoo Kim, Michael Laskin, Igor Mordatch, and Deepak Pathak. 2021. How to adapt your large-scale vision-and-language model. (2021).
- [39] Romeo Lanzino, Federico Fontana, Anxhelo Diko, Marco Raoul Marini, and Luigi Cinque. 2024. Faster than lies: Real-time deepfake detection using binary neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3771–3780.
- [40] Binh Le, Shahroz Tariq, Alsharif Abuadba, Kristen Moore, and Simon Woo. 2023. Why do facial deepfake detectors fail?. In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*. 24–28.
- [41] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollar. 2014. Microsoft COCO: Common objects in context. *European conference on computer vision* (2014), 740–755.
- [42] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. 2023. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272* (2023).
- [43] Author: Tailor made AI. 2019, November. *Yijian: Chinese AI Painting Creative Cloud*. <https://creator.nightcafe.studio/> [Accessed on 12-04-2025].

- [44] Tailor made AI. 2022. *Dreambooth: Tailor-made AI Image Generation*. <https://www.astria.ai/> [Accessed on 12-04-2025].
- [45] Sara Mandelli, Davide Cozzolino, Edoardo D Cannas, Joao P Cardenuto, Daniel Moreira, Paolo Bestagini, Walter J Scheirer, Anderson Rocha, Luisa Verdoliva, Stefano Tubaro, et al. 2022. Forensic analysis of synthetically generated western blot images. *IEEE Access* 10 (2022), 59919–59932.
- [46] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)* 54, 1 (2021), 1–41.
- [47] Emad Mostaque. 2022. *Stabilityai: Stable Diffusion Public Release*. <https://stability.ai/blog/stable-diffusion-public-release> [Accessed on 12-04-2025].
- [48] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Du Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. 2022. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding* 223 (2022), 103525.
- [49] Fan Nie, Jiangqiu Ni, Jian Zhang, Bin Zhang, and Weizhe Zhang. 2024. FRADE: Forgery-aware Audio-distilled Multimodal Learning for Deepfake Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6297–6306.
- [50] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.
- [51] Or Patashnik, Rinon Gal, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *arXiv:2108.00946 [cs.CV]*
- [52] Alvaro Lopez Pellicer, Yi Li, and Plamen Angelov. 2024. PUDD: towards robust multi-modal prototype-based deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3809–3817.
- [53] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
- [54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [55] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Alec Radford, Mark Chen, and Ilya Sutskever. 2022. DALL-E 2. *arXiv:2201.03994 [cs.CV]*
- [56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- [57] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.
- [58] Angus Russell. 2019, November. *Nightcafe: Create Amazing Artworks using the Power of Artificial Intelligence*. <https://creator.nightcafe.studio/> [Accessed on 12-04-2025].
- [59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.
- [60] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28130–28139.
- [61] Jianhua Tao, Jiangyan Yi, Cunhang Fan, Ruibo Fu, Shan Liang, Pengyuan Zhang, Haizhou Li, Helen Meng, Dong Yu, and Masato Akagi. 2022. DDAM’22: 1st International Workshop on Deepfake Detection for Audio Multimedia. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7405–7406.
- [62] Chenyang Wang, Xuehai Liu, Han Zhang, Yixin Wang, Yiming Sheng, Lu Yuan, Jun-Yan Zhu, and Lujia Yang. 2023. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2302.08113* (2023).
- [63] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695–8704.
- [64] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22445–22455.
- [65] Taiba Majid Wani, Reeva Gulzar, and Irene Amerini. 2024. Abc-capsnet: Attention based cascaded capsule network for audio deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2464–2472.
- [66] Haiwei Wu, Jiantao Zhou, and Shile Zhang. 2023. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800* (2023).
- [67] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. 2024. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8984–8994.
- [68] Daichi Zhang, Fanzhao Lin, Yingying Hua, Pengju Wang, Dan Zeng, and Shiming Ge. 2022. Deepfake video detection with spatiotemporal dropout transformer. In *Proceedings of the 30th ACM international conference on multimedia*, 5833–5841.
- [69] Junbin Zhang, Yixiao Wang, Hamid Reza Tohidpour, and Panos Nasiopoulos. 2023. Detecting stable diffusion generated images using frequency artifacts: A case study on disney-style art. In *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1845–1849.
- [70] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. 2019. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 1–6.
- [71] Yi Zhang, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Weibin Yao, Zhe Li, Bingyu Hu, Weiwei Feng, Tao Gong, et al. 2024. MFMS: Learning Modality-Fused and Modality-Specific Features for Deepfake Detection and Localization Tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11365–11369.
- [72] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.
- [73] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems* 36 (2023), 77771–77782.
- [74] Tianlei Zhu, Junqi Chen, Renzhe Zhu, and Gaurav Gupta. 2023. StyleGAN3: generative networks for improving the equivariance of translation and rotation. *arXiv preprint arXiv:2307.03898* (2023).

(Supplementary Materials)

SpecXNet: A Dual-Domain Convolutional Network for Robust Deepfake Detection

Inzamamul Alam

Department of Computer Science and
Engineering

Sungkyunkwan University
Suwon, Republic of Korea
inzi15@g.skku.edu

Md Tanvir Islam

Department of Computer Science and
Engineering

Sungkyunkwan University
Suwon, Republic of Korea
tanvirnwu@g.skku.edu

Simon S. Woo*

Department of Computer Science and
Engineering

Sungkyunkwan University
Suwon, Republic of Korea
swoo@g.skku.edu

a) Original/Fake Images



b) Fast Fourier Transformed Results

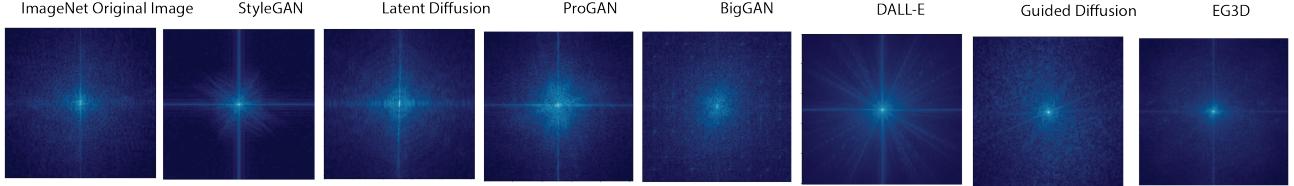


Figure 1: Examples of the AI-generated fake images and their corresponding FFT results: (Top) images generated by state-of-the-art generative models: StyleGAN3 [10], Latent Diffusion [4], ProGAN [6], BigGAN [1], DALL-E [8], Guided Diffusion [3], and EG3D [2], respectively. The bottom images have corresponding FFT results that have specific frequencies corresponding to each generated model.

1 FFT SIGNATURES OF GENERATIVE MODELS

Figure 1 showcases visual comparisons between original ImageNet images and outputs from various state-of-the-art generative models, including StyleGAN, Latent Diffusion, ProGAN, BigGAN, DALL-E, Guided Diffusion, and EG3D. The top row presents synthesized samples from each generator, while the bottom row illustrates the corresponding 2D Fast Fourier Transforms (FFT) of these images.

Notably, each generative model produces a distinct frequency signature in the Fourier domain. While natural images exhibit dense but smoothly decaying high-frequency spectra, AI-generated samples reveal unique structural patterns—such as radial lines, symmetry artifacts, and spectral voids—arising from upsampling heuristics, architectural inductive biases, or latent priors.

These frequency-domain discrepancies provide a powerful modality for manipulation detection, motivating our dual-domain architecture. The Spectral Fusion framework explicitly models such

artifacts by decomposing input features into spatial and spectral branches, allowing our model to attend to both texture-level anomalies and frequency-based inconsistencies. As highlighted in this figure, leveraging FFT-based representations reveals generator-specific spectral traces that are often imperceptible in the spatial domain.

2 ADDITIONAL INFORMATION ON DATASET

We have used many prompts to generate images using stable diffusion V1.4 [9], midjourney V5 [5], Dreambooth [7], etc. Here are some samples of the prompts that we used for image generation.

- (1) Bruce Lee sitting in a car on a road way.
- (2) Mother Teresa portrait from front, she is serious with a beautiful makeup.
- (3) Cyberpunk sci-fi woman portrait sitting in a pipe light session.
- (4) An angry 6 year old girl steering in jungle cloths having burn marks on her face.

*Corresponding author. Email: swoo@g.skku.edu (Simon S. Woo)

Table 1: Complexity analysis of SpecXNet. We report parameter and FLOPs estimates for the vanilla Xception block, the proposed SpecXNet components—including the Dual-Domain Feature Coupler (DDFC), Spectral Transform, and Dual Fourier Attention (DFA)—as well as the total per-block cost. C_1, C_2 are input/output channels, $H \times W$ is spatial resolution, K is kernel size, and α determines the global channel ratio.

Module	Parameter Count	FLOPs Estimate
Baseline (XceptionNet)	$C_1 C_2 K^2$	$C_1 C_2 K^2 H W$
DDFC		
Local → Local	$(1 - \alpha)^2 C_1 C_2 K^2$	$(1 - \alpha)^2 C_1 C_2 K^2 H W$
Global → Global (Spectral)	$\alpha^2 C_2 (\frac{1}{2} C_1 + \frac{3}{2} C_2)$	$\alpha^2 C_2 H W (\frac{1}{2} C_1 + \frac{13}{16} C_2)$
Local → Global	$\alpha(1 - \alpha) C_1 C_2 K^2$	$\alpha(1 - \alpha) C_1 C_2 K^2 H W$
Global → Local	$\alpha(1 - \alpha) C_1 C_2 K^2$	$\alpha(1 - \alpha) C_1 C_2 K^2 H W$
DFA	$2(C_1 + C_2)^2$	$2(C_1 + C_2)^2 + (C_1 + C_2) H W$
Total (SpecXNet Block)	$(1 - \alpha^2) C_1 C_2 K^2 + \alpha^2 C_2 (\frac{1}{2} C_1 + \frac{3}{2} C_2) + 2(C_1 + C_2)^2$	$(1 - \alpha^2) C_1 C_2 K^2 H W + \alpha^2 C_2 H W (\frac{1}{2} C_1 + \frac{13}{16} C_2) + 2(C_1 + C_2)^2 + (C_1 + C_2) H W$

Table 2: Efficiency comparison of SpecXNet and ResNet152+UGAD on RTX 3090 and GTX 1660 Ti. SpecXNet achieves significantly higher throughput, lower latency, and reduced VRAM usage across both platforms.

Method	IPS ↑ (RTX 3090)	IT ↓ (RTX 3090)	VRAM (GB) (RTX 3090)	IPS ↑ (GTX 1660 Ti)	IT ↓ (GTX 1660 Ti)	VRAM (GB) (GTX 1660 Ti)
ResNet152 + UGAD	46	21.6	2.5	12	84.3	2.3
SpecXNet (ours)	112	6.3	1.2	33	23.3	1.1

- (5) Visualize a mountain top villa with trees growing on top of it. A man is standing in the building to watch the beautiful landscape.
- (6) An old kitten portrait in a woman cold weather dress.
- (7) Generate an image of a bustling space station with diverse inhabitants and advanced technology, inspired by the text.
- (8) Realistic, hyper resolution, joker turned into vampire, laughing portrait.
- (9) A canal passing through jungle covered with trees.
- (10) A dog Pirate cartoon standing with his cap sailing through the sea.
- (11) Pikachu standing on a rock on sea side.
- (12) A Sikh bodybuilder handsome beard man portrait.
- (13) Anime-style Japanese woman eating super in
- (14) A front portrait of an Anime-style Japanese boy crying in his Japanese room, his face is red from his nose and cheeks.
- (15) A street cat walking ahead on the cyberpunk street.

3 COMPUTATIONAL EFFICIENCY

Although our proposed methodology introduces an intricate dual-domain representation framework and adaptive attention mechanisms, its computational complexity remains highly manageable, largely due to strategic algorithmic choices and efficient spectral-domain operations. Specifically, the global spectral branch capitalizes extensively on the computational efficiency of the Fast Fourier Transform (FFT), a cornerstone algorithm whose complexity is well-studied and thoroughly optimized in numerical libraries.

Formally, the computational complexity for a two-dimensional FFT on an input feature map of dimensions $H \times W$ with αC spectral channels can be analytically expressed as follows:

$$O(\alpha C \cdot H \cdot W \cdot \log(HW)), \quad (1)$$

where α denotes the proportion of channels allocated specifically to spectral domain operations. Importantly, the $\log(HW)$ scaling factor underscores a significant efficiency advantage relative to traditional spatial convolutions that typically exhibit quadratic computational complexity relative to the kernel size. Thus, even with increasing spatial resolution or deeper layers that traditionally impose computational strain, the FFT-based spectral processing retains high efficiency.

To precisely quantify and underscore these efficiency gains, consider a traditional convolutional operation characterized by kernel size $k \times k$ across αC input channels producing similar global receptive fields. The conventional convolution complexity scales quadratically as:

$$O(\alpha C \cdot k^2 \cdot H \cdot W), \quad (2)$$

which quickly becomes computationally prohibitive as receptive fields and kernel sizes increase to capture large-scale global contexts. In stark contrast, the spectral approach, through its frequency-domain computation, effectively sidesteps this quadratic scaling by transforming spatial convolutions into efficient element-wise multiplications in frequency space. The complexity reduction achieved by this spectral method is profound, especially for large spatial dimensions or extensive receptive fields, typically required to capture global information.

Further enhancing efficiency, the spectral modulation step—where frequency coefficients are adaptively weighted—employs lightweight operations. Specifically, spectral modulation involves learned element-wise multiplicative operations on frequency components, amounting to a complexity of:

$$O(\alpha C \cdot H \cdot W), \quad (3)$$

significantly lower than spatial convolutions. Additionally, the inverse FFT that maps modulated global spectral features back to

the spatial domain incurs computational complexity similar to the forward FFT, preserving the overall complexity advantage:

$$\mathcal{O}(\alpha C \cdot H \cdot W \cdot \log(HW)). \quad (4)$$

Moreover, our Dual Fourier Attention (DFA) mechanism introduces minimal computational overhead by employing efficient global pooling and compact linear transformations. Global average pooling (GAP), used to summarize spatial feature maps, scales linearly with spatial resolution:

$$\mathcal{O}(C \cdot H \cdot W), \quad (5)$$

and subsequent linear transformations for attention calculation scale merely linearly with channel dimensionality:

$$\mathcal{O}(C^2), \quad (6)$$

which is negligible relative to convolutional layers in deep architectures.

These theoretical estimates are further validated by the empirical accounting provided in Table 1, which details the parameter and FLOP costs for each dual-domain component in SpecXNet. The local-to-local and cross-branch convolutional paths scale quadratically with the kernel size, as in traditional CNNs. In contrast, the spectral transform path replaces large-kernel spatial convolutions with FFT/iFFT operations and frequency-domain modulation, achieving favorable $\mathcal{O}(\log HW)$ scaling relative to spatial size.

Importantly, the table captures practical design elements such as the dual-branch decomposition in the DDFC, the cost of spectral modulation, and the full complexity of the Dual Fourier Attention (DFA) module. Despite the added modeling capacity, the DFA introduces only a minor overhead ($\mathcal{O}((C_1 + C_2)^2)$), as it relies on global average pooling and lightweight linear projections. When aggregated, the full SpecXNet block maintains computational efficiency competitive with the vanilla XceptionNet baseline while offering significantly enhanced representational power.

Furthermore, we conducted experiments for a runtime comparison between SpecXNet and ResNet152+UGAD on both RTX 3090 and GTX 1660 Ti, as presented in Table 2. SpecXNet achieves higher throughput, lower inference time, and reduced memory usage on both platforms, demonstrating its efficiency and suitability for real-time applications.

Table 3: Accuracy across different values of α , controlling the spectral-to-spatial channel ratio in DDFC.

α	0.0 (all spatial)	0.25	0.50	0.75
Acc (%)	87.9	92.7	96.3	93.2

4 ABLATION STUDY ON α RATIO IN DDFC

We performed an ablation study on the spectral-to-spatial channel ratio α in the Dual-Domain Feature Coupler (DDFC). As shown in Table 3, setting $\alpha = 0.5$ provided the highest accuracy, suggesting that a balanced division between the spatial and spectral pathways offers the most effective feature representation. Fully spatial or heavily unbalanced configurations result in lower performance.

5 CODE AVAILABILITY

The implementation of the proposed method will be made publicly available upon acceptance of the paper. This will include training scripts, model checkpoints, and evaluation code to facilitate reproducibility and further research.

REFERENCES

- [1] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).
- [2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16123–16133.
- [3] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- [4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [5] David Holz. 2022, July 12. *Midjourney: Expanding the Imaginative Powers*. <https://www.midjourney.com/home> [Accessed on 12-04-2025].
- [6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).
- [7] Tailor made AI. 2022. *Dreambooth: Tailor-made AI Image Generation*. <https://www.astria.ai/> [Accessed on 12-04-2025].
- [8] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Alec Radford, Mark Chen, and Ilya Sutskever. 2022. DALL-E 2. *arXiv:2201.03994 [cs.CV]*
- [9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [10] Tianlei Zhu, Junqi Chen, Renzhe Zhu, and Gaurav Gupta. 2023. StyleGAN3: generative networks for improving the equivariance of translation and rotation. *arXiv preprint arXiv:2307.03898* (2023).