Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multi-Domain Feature Fusion Framework for Generalizable Deepfake Detection Across Different Generators

Amna Amjid
National University of Sciences and Technology
Sana Qadir
National University of Sciences and Technology
Mehwish Fatima
National University of Sciences and Technology
Raja Khurram Shahzad
raja-khurram.shahzad@miun.se

Mid Sweden University

Article

Keywords: Deepfake Detection, Generalization, Multi-Domain Learning

Posted Date: October 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11104862/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multi-Domain Feature Fusion Framework for Generalizable Deepfake Detection Across Different Generators

Amna Amjid[1], Sana Qadir[1*], Mehwish Fatima[1], Raja Khurram Shahzad[2,3*]

[1*]School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan.
[2]Department of Communication, Quality Management and Information Systems, Mid Sweden University, Ostersund Campus, Sweden.
[3]Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Luleå, Sweden.

*Corresponding author(s). E-mail(s): sana.qadir@seecs.edu.pk; Raja-Khurram.Shahzad@miun.se;

Abstract

Deepfakes are artificially generated images, audio, or videos that threaten privacy, security, and information integrity. Detecting such content has become increasingly challenging as recent generative models produce highly realistic synthetic media with substantially different generation characteristics. Existing deepfake image detectors commonly rely on spatial, gradient, or frequency-domain artifacts in isolation, making their learned representations sensitive to generator-specific distributions and limiting their transferability to unseen generative models. This paper proposes the Spatial-Gradient-Frequency Fusion Network (SGFF-Net), a multi-domain deepfake image detection framework that integrates spatial, gradient, and Discrete Wavelet Transform (DWT)-based frequency representations within a dual residual learning architecture. Rather than evaluating detection only under conventional intra-dataset settings, SGFF-Net is systematically assessed under cross-model, cross-paradigm, multi-source, and real-world distribution shifts. The results indicate that multi-domain feature fusion provides more transferable forensic representations than individual spatial, gradient, or frequency-based representations, although the degree of transfer depends on the source and target generator distributions.

Cross-paradigm experiments further highlight an asymmetric transfer between GAN- and diffusion-based generators, demonstrating that robustness across one generator family does not necessarily imply robustness across fundamentally different synthesis paradigms. Data augmentation primarily improves transfer under larger distribution shifts, whereas heterogeneous multi-source training reduces source-specific dependence by exposing the detector to more diverse generative characteristics. Evaluation on independently collected real-world images nevertheless reveals a substantial remaining domain gap. These findings demonstrate that generalizable deepfake detection relies on a combination of complementary forensic representations, generators' diversity, and training variability, while robustness to previously unseen real-world generation processes remains an open challenge.

Keywords: Deepfake Detection, Generalization, Multi-Domain Learning.

1 Introduction

Recent advances in Generative Artificial Intelligence (GenAI) have made it increasingly easy to create highly realistic synthetic media, commonly referred to as *deepfakes*. These deepfakes can include images, videos, and audio. As a result, they can now be produced on a large scale with minimal technical expertise, leading to their growing use in disinformation campaigns, impersonation attacks, fraud, blackmail, and privacy violations [1–3]. This rising threat has transformed deepfakes from a mere technological issue into a significant societal and security challenge. The *Global Risks Report 2026* ranks misinformation and disinformation as the fifth most critical global risk, predicting they will continue to have a substantial impact in the coming years [4].

Deepfake detection has become increasingly challenging with the emergence of advanced generative models, such as Stable Diffusion (SD) [5], DALL-E V [2], and StyleCLIP [6]. These models produce highly realistic synthetic content and significantly enhance the variety of generation methods and the resulting artifacts. In addition to enabling misinformation and fraud, deepfakes pose a growing threat to biometric authentication systems [1].

This study specifically focuses on image deepfake detection, as manipulated images are widely disseminated across online platforms and pose significant forensic and security challenges. Existing image deepfake detection methods broadly fall into two categories. Cue-based approaches exploit artifacts introduced during content generation, including spatial irregularities, texture inconsistencies, and frequency-domain anomalies [7–12]. Because these artifacts are often interpretable, cue-based methods provide forensic evidence that can support detection decisions. In contrast, data-driven approaches train Machine Learning (ML) and Deep Learning (DL) models to learn discriminative representations directly from data [2, 13–16].

Early detection methods primarily focused on visible spatial artifacts, such as abnormal textures, lighting inconsistencies, and unrealistic facial structures [7–9, 17, 18]. Subsequent research shifted to frequency-domain analysis, which exploits spectral

patterns introduced during image synthesis [10–12, 19–21]. More recently, gradient-based approaches have demonstrated the ability to capture structural inconsistencies and generator-induced traces that are less apparent in the pixel domain [22, 23].

However, these representations exhibit different sensitivities to the underlying generation process. Spatial artifacts may vary with image content and synthesis quality, while frequency and gradient artifacts often depend on the specific architecture and generation procedure used. Therefore, strong detection performance within one generator distribution does not necessarily translate to unseen generators or fundamentally different generation paradigms. This distinction is particularly important when comparing GAN-based synthesis to diffusion-based synthesis, as the mechanisms underlying these two approaches differ significantly [2, 24].

Although CNN-based detectors perform well in controlled environments, their ability to learn transferable forensic representations across unseen generators remains limited [13, 14]. To improve robustness, several studies have incorporated frequency-domain representations, including Fast Fourier Transform (FFT) [25, 26], Discrete Cosine Transform (DCT) [27], and Discrete Wavelet Transform (DWT) [28–30]. These transformations expose synthesis artifacts that are often less apparent in the spatial domain and have been shown to improve detection performance and cross-dataset robustness. Nevertheless, existing research primarily focuses on developing and evaluating frequency-aware detectors on GAN-generated datasets, which often exhibit degraded performance when applied to diffusion-generated content, and vice versa. Despite the growing use of diffusion models, comparatively limited attention has been given to forensic representations that transfer across both GAN and diffusion paradigms.

The challenge is therefore not only to identify discriminative artifacts, but also to determine whether complementary representations can provide forensic cues that remain useful as the generator changes. Recent surveys suggest that improving robustness requires moving beyond single-domain representations and leveraging complementary information from multiple feature spaces [2]. Motivated by this observation, we propose a multi-domain deepfake detection framework that integrates spatial, gradient, and frequency-domain representations within a Dual Residual Network (DRN) architecture. The proposed Spatial-Gradient-Frequency Fusion Network (SGFF-Net) combines RGB information with gradient-based artifacts and localized multi-resolution frequency information extracted using DWT. Rather than evaluating the framework only under conventional intra-dataset conditions, we examine how these representations perform under progressively different distributions, including unseen generators, cross-paradigm transfer, heterogeneous multi-source training, and independently collected real-world images.

The main contributions of this paper are summarized as follows:

- We propose SGFF-Net, a multi-domain deepfake detection framework that simultaneously leverages spatial, gradient, and DWT-based frequency representations within a novel, unified dual residual learning architecture. The architectural choices, including the frequency representation, DWT configuration, and feature-branch combination, are empirically established through systematic ablation studies.

- We conduct a structured evaluation of SGFF-Net across intra-dataset, cross-model, cross-paradigm, multi-source, and real-world settings. This evaluation aims to investigate how multi-domain forensic representations transfer as the gap between training and evaluation distributions increases. The experiments reveal an asymmetric transfer between GAN-based and diffusion-based generators, indicating that generalization across different generative paradigms remains dependent on the source and target distributions.
- We systematically investigate the impact of training-data diversity and augmentation on generalization. The findings suggest that heterogeneous multi-source training reduces dependence on source-specific data, while augmentation proves to be most beneficial under challenging distribution shifts. Moreover, evaluations on independently collected real-world images further demonstrate that combining feature diversity with training-data diversity improves transferability, although a substantial real-world domain gap remains.

The remainder of this paper is structured as follows. Section 2 reviews related work on deepfake detection, including cue-based, data-driven, and diffusion-based detection methods. Section 3 describes the proposed framework. Section 4 presents details of the experiments and implementation. Section 5 provides the results, while Section 6 discusses these findings. Finally, the conclusion is given in Section 7.

2 Related Work

This section discusses different approaches for deepfake image detection, including cue-based and data-driven approaches. Cue-based detection methods exploit spatial, gradient-domain, or frequency artifacts introduced by deepfake generators as discriminative forensic cues. Additionally, recent deepfake detection approaches for DM-generated media are also reviewed in this section.

2.1 Spatial-Based Deepfake-Image Detection

To mitigate the threats posed by sophisticated deepfakes, researchers have investigated spatial-domain forensic cues. Early studies used color-space inconsistencies [7–9] and natural texture irregularities [17, 18] as discriminative indicators of synthetic images. Yu et al. [31] and Marra et al. [32] suggested that every GAN-generated image contains hidden generator-specific fingerprints in the spatial domain that can serve as reliable indicators for the digital media forensic community. Their work further investigated the stability, robustness, and source specificity of these fingerprints across different GAN architectures. During experimentation, the data indicated that even small variations in training create distinct fingerprints, suggesting that these signatures are associated with generator-specific training dynamics rather than merely visible image artifacts. However, the studies assumed that GAN classes are known during training/testing, resulting in poor open-set generalization. As a result, these methods are effective when the source generator is represented during training, but exhibit limited generalization to previously unseen generators and emerging synthesis paradigms.

Tan et al. [22] identified this limitation when they stated that there is no sufficiently generalized artifact representation for GAN-generated image detection. They hypothesized that the gradient domain of deepfake-generated images might contain discriminative regions, high-level artifacts, and generalized clues that are better suited for unseen Generative Adversarial Networks (GANs). Based on this, they proposed LGrad (Learning on Gradients), which passes each image through a fixed pre-trained transformation model and computes input gradients as the representation for subsequent classification. The gradient representation is intended to emphasize generator-induced artifacts while suppressing semantic image content. Although LGrad reported an average improvement of 11.4% over previous detectors, its evaluation remained largely centered on GAN-generated images. Consequently, its ability to generalize to face-swapping methods and diffusion-based generators remained limited, indicating that gradient representations alone are insufficient for universal deepfake detection.

2.2 Frequency-Based Deepfake-Image Detection

In the frequency domain, researchers investigate artifacts introduced by generative models, such as inconsistent high-frequency textures or unnatural smoothness in low-frequency regions [10, 11, 19]. Frequency transforms such as FFT, DCT, and DWT are used to expose these artifacts and assess their usefulness as transferable forensic cues. Existing methods, e.g., the multi-feature channel domain-weighted (MCW) meta-learning framework [27], spatial-phase shallow learning (SPSL) [10], frequency-enhanced self-blended images (FSBI) [28], and Frequency-aware Discriminative Feature Learning framework (FDFL) [20], exploit frequency extraction methods to capture abnormalities in the frequency domain. These abnormalities are then used as detection cues. Uddin et al. proposed a deepfake face detection approach that combines the outputs of a multi-level DWT and a vision transformer to obtain frequency features, distinguishable information, and global forgery cues [29]. Giudice et al. analytically demonstrated the nature of multimedia content as real or deepfake using DCT to detect so-called GAN-specific frequencies (GSFs) [21]. Their method models the 63 Alternating Current (AC) coefficients, corresponding to the frequency-detail components of 8×8 DCT blocks, with a Laplacian distribution and extracts the β scale parameter, which measures the spread or variation of each frequency component. The results demonstrate that the β statistics inferred from the distribution of AC coefficients are key to distinguishing GAN-generated data and serve as explainable fingerprints of GAN-generated images.

More recently, Sharafudeen and Chandra proposed a Dual Residual Network (DRN) for frequency-based deepfake detection [25]. Their framework applies FFT-based preprocessing and uses residual learning to capture local and global frequency-domain forensic patterns. However, the evaluation primarily focused on intra-distribution settings, and the study did not investigate robustness across unseen generator families or diffusion-based models. To detect both subtle and advanced forgery traces, Qian et. al. proposed a Frequency in Face Forgery Network (F^3-Net) that utilizes a dual-path architecture designed to reveal manipulation footprints

Table 1 Comparison of representative deepfake and synthetic image detection methods in terms of feature domain, methodology, key findings, and limitations

Ref.	Feature / Domain	Methodology	Key Findings	Limitations
[22]	Spatial / Gradient	LGrad using pre-trained CNN, gradient representation, and classifier	Gradient artifacts improved cross-model generalization by up to 11.4%	Weak on non-GAN deepfakes; sensitive to JPEG compression and noise
[33]	Data-driven	CNN + Multi-Head Self-Attention (MHSA) with attention heat-maps	Achieved 94–98% accuracy and up to 99.75% AUC using local and global features	Poor generalization on unseen datasets and generators
[13]	Data-driven	Siamese Network with contrastive learning and CNN classifier (DeepFeatureX-SN)	Achieved 97.29% accuracy and improved cross-generator performance (67.40%)	Limited robustness on unseen generators; high computational cost
[34]	Spatial + Semantic	Frozen CLIP ViT-L/14 feature extractor with Nearest Neighbor and Linear Probe classifiers	Achieved strong universal generalization across GAN, diffusion, and autoregressive image generators	Relies on large-scale CLIP pre-training; high computational and memory requirements
[25]	Frequency	FFT-based frequency forensics with Dual Residual Network (DRN)	Residual learning captured frequency fingerprints and improved detection performance	Strong dataset bias and limited leave-one-generator-out generalization
[28]	Frequency	Self-Blended Images (SBI) augmentation with DWT-based features	DWT exposed subtle forgery artifacts and improved cross-dataset performance	Limited cross-generator and diffusion-model evaluation
[35]	Frequency	CNN-based frequency analysis for GAN and diffusion image detection	Frequency artifacts provided strong discriminative forensic cues	Poor generalization to unseen generators and datasets
[36]	Spatial + Semantic	Multimodal diffusion attribution using image and text-prompt embeddings	Captured diffusion-specific fingerprints using semantic cues	Closed-set limitation; requires prompts and lacks real-world robustness
[26]	Frequency	FFT high-pass filtering with spectral peak extraction and gradient boosting	Effective on multiple diffusion models with partial compression robustness	Sensitive to domain shift and unseen diffusion models

through the collaborative analysis of decomposed frequency features and local probabilities [19]. Jeong et al. proposed a robust deepfake detection system called BiHPF consisting of two high-pass filters (HPFs) [12]. One filter is a frequency-level HPF that amplifies the magnitude, and the second is a pixel-level HPF that emphasizes the surrounding background pixel values. These two HPFs jointly increase the visibility of discriminator artifacts. In another similar work, Jeong et al. observed that the generalization capability of detection models weakens because frequency artifacts in generated images are highly sensitive to the generator's distribution and the object categories [11]. They used this insight to develop FrePGAN, a system that improves model generalization by introducing frequency-level perturbation maps to remove the effect of the frequency-level artifacts.

Although spatial, gradient, and frequency representations have individually been investigated (e.g., DRN [25], FSBI[28], and MCW [27]) and successfully demonstrated that transformed representations can expose forensic cues from RGB images, these studies primarily utilise a particular representation or frequency transformation. Moreover, the discriminative capability of these methods varies substantially across datasets and generative paradigms. This leaves an open question of whether fusing complementary spatial, gradient, and frequency information can improve robustness across different generative paradigms.

2.3 Data-driven Deepfake-Image Detection

Although studies that focus only on the frequency domain show satisfactory performance for intra-dataset evaluation on GAN-based datasets, performance degrades drastically in cross-dataset settings (especially with recent diffusion models), indicating poor generalizability. Another issue with GAN-generated faces is that it is possible to remove GAN fingerprints from the frequency domain using adversarial techniques, thereby disabling detectors based on such cues [2, 33]. To overcome such limitations, research has shifted towards data-driven deepfake detection rather than manually extracting handcrafted cues from the spatial or frequency domains. Data-driven deepfake detection utilizes Deep Neural Networks (DNNs) and large amounts of training data, rather than relying on specific feature extraction methods to extract Principal Discriminative Features (PDFs). For instance, Marra et al. conducted a comparative study of deepfake detection using images generated by multiple GAN architectures, including CycleGAN[37, 38]. Their analysis indicated that XceptionNet outperformed the other evaluated networks; however, the study did not investigate cross-generator generalization.

Ameer et al. investigated whether data augmentation genuinely improves the deepfake detection rate and generalizability of data-driven models [39]. They compared three different CNN models (XceptionNet, ResNet-50, and EfficientNet-B3) and two transformer models (ViT-B/16 and ConvNeXt-Small), both with and without augmentation, using the GAN-based Deepfake Detection Challenge (DFDC) dataset [40]. Their cross-model experimental results for XceptionNet (89%) and ViT-B/16 (75%) indicated that data augmentation improves cross-model generalization more in CNNs than in transformer models.

Recently, Pontorno et al. used data-driven self-supervised contrastive learning to propose a DeepFeatureX-SN framework [13]. Architecturally, it leveraged tripartite Siamese DenseNet161 extractors operating in parallel to mitigate feature-extraction challenges. Each extractor was tasked with deriving convolutional features from only one image type: real, GAN, or DM-generated. The evaluation of the framework exhibited strong performance by achieving 97.70% accuracy for binary classification (for the same dataset evaluation). During the generalization assessment, the framework achieved 74.45% accuracy on unseen GAN-generated images and 57.24% on unseen DM-generated images. While DeepFeatureX-SN outperformed existing state-of-the-art methods, it had high computational requirements and relatively low performance on unseen DM-generated images.

2.4 Diffusion Model Deepfake-Image Detection

Recent years have seen increasing adoption of diffusion-based image synthesis, accompanied by substantial improvements in image quality and realism. Representative models include Denoising Diffusion Implicit Models (DDIM) [41], Denoising Diffusion Probabilistic Models (DDPM) [42], SD [5], Midjourney, DALL·E 3 [43]. These models enable real-time text-to-image generation using only prompts, further lowering the barrier to using deepfakes. However, most existing deepfake detection methods focus largely on distinguishing real content from GAN-generated content. In this section, the

few notable studies that specifically tackle the detection of deepfake images generated by DMs are discussed.

In 2023, Corvi et al. investigated the problem by attempting to understand how difficult it is to distinguish synthetic images generated by diffusion models from real ones and whether current SOTA detectors were suitable for the task [35]. They retrained the existing architecture of Gragnaniello et al. [44] on DM-generated images and used the findings to demonstrate that synthetic images generated by GANs and DMs have specific fingerprints that depend on the architecture and parameters of the generative network used to produce them.

In the same year, Sha et al. proposed DE-FAKE, a method based on an ML classifier for diffusion model detection on four popular text-to-image architectures (including DALLE 2 and Latent Diffusion) [36]. Their empirical results indicated that deepfake images generated by different diffusion models can be distinguished from real ones using frequency-domain analysis due to common artifacts in diffusion-generated images. Additionally, they demonstrated that different diffusion models leave unique fingerprints in their generated images, enabling deepfake images to be effectively attributed to their source models.

Ojha et al. proposed UniversalFakeDetect (UFD), an open-world synthetic image detector based on representations obtained from a pretrained CLIP encoder [34]. Using nearest-neighbour classification and linear probing, they demonstrated improved transfer to unseen GAN, diffusion, and autoregressive generators.

Bammey investigated synthetic image detection and proposed a spectral analysis-based method for diffusion models [26]. They extracted artifacts of residual images left during the Fourier transformation in the diffusion process and used manually selected frequency components for detection. A key contribution of the work is the demonstration that a simple cross-difference high-pass filter can reveal diffusion-related frequency artifacts more effectively than prior residual extraction approaches, thereby enabling their detection at the individual-image level.

Chen et al. developed a masked conditional DM for deepfake detection and specifically studied the use of augmentation to improve generalization [45]. Instead of directly swapping faces and using traditional augmentations (flip, rotation, blur, etc.), they randomly masked facial regions and used a conditional diffusion model to reconstruct or manipulate the masked areas, generating a diverse range of forged faces with subtle artifacts. During experimentation, the data showed clear improvement in cross-dataset generalization. For example, the AUC for the Celeb-DF dataset improved from 73.12% (without augmentation) to 78.02% using their diffusion-based augmentation.

Table 1 presents a summary of notable image deepfake detection methods discussed above in terms of feature representation, methodology, key findings, and limitations. The reviewed literature indicates that individual spatial, gradient, frequency, and semantic representations exhibit different generalization characteristics across generative models. Most existing detectors nevertheless rely predominantly on one representation domain, limiting their ability to exploit complementary forensic information when the generator distribution changes. A few hybrid approaches have been developed, such as DE-FAKE, which improve detection by incorporating prompt semantics, but they depend on auxiliary textual information that might not always

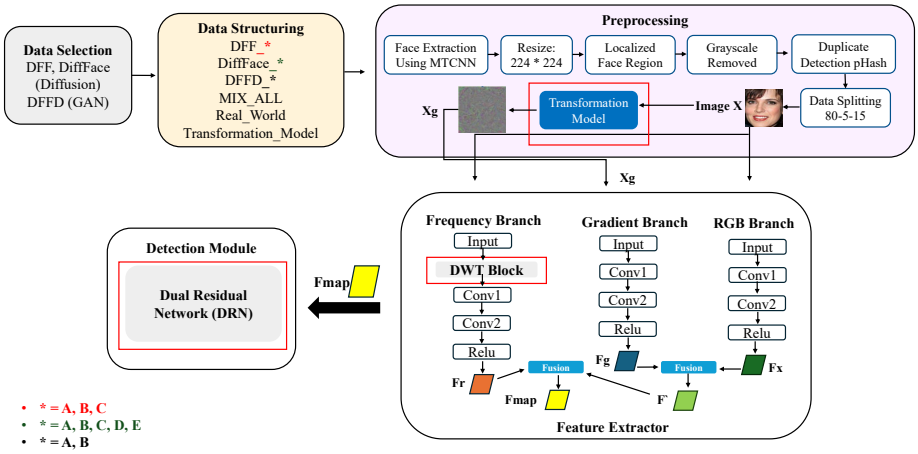


Fig. 1 Overview of Spatial-Gradient-Frequency Fusion Network (SGFF-Net): the proposed multi-domain deepfake detection framework. RGB, gradient, and DWT-based frequency representations are extracted in parallel and fused before classification using the Dual Residual Network (DRN).

be available in real-world scenarios [36]. Additionally, most existing frequency analyses are limited, focusing primarily on analyzing high-frequency artifacts and spectral peaks [25, 26]. In other words, they lack localized multi-resolution frequency learning. Also, recent studies have emphasized the importance of robust feature representations and augmentation strategies for improving generalization, yet these techniques remain underexplored [2, 46]. Furthermore, many existing methods demonstrate strong intra-dataset performance but suffer considerable performance degradation during cross-dataset and cross-generator evaluations [22, 25]. This paper aims to address the above-mentioned research gaps by achieving the following two research objectives:

1. Design and develop a multi-domain deepfake detection framework called SGFF-Net that integrates spatial (RGB), gradient, and DWT-based frequency representations within a Dual Residual Learning (DRN) architecture for improved feature representation and generalization.
2. Evaluate the generalization of SGFF-Net in cross-model, cross-paradigm, multi-source, and real-world scenarios (including with and without augmentation and diverse training strategies).

3 Proposed Framework: SGFF-Net

Figure 1 illustrates the architecture of our proposed framework, Spatial-Gradient-Frequency Fusion Network (SGFF-Net) for Generalizable Deepfake Detection. It consists of four stages: data selection and structuring, preprocessing, feature extraction, and DRN-based detection. The framework begins by selecting three publicly available datasets: DeepFakeFace (DFF) [47] Diffusion Face (DiffFace) [48] and Diverse

Fake Face Dataset (DFFD) [40] (GAN-based), and uses their images to create multiple datasets for systematic evaluation. The preprocessing model processes all images (including generated ones) and performs face extraction and localization using Multi-task Cascaded Convolutional Networks (MTCNN). Images for which face detection was unsuccessful were discarded. The resulting preprocessed images were then divided into training, validation, and testing subsets using an 80:5:15 split. To minimize class imbalance, each experimental dataset was designed with approximately equal numbers of real and fake image instances. To prevent identity leakage, datasets were partitioned at the identity level rather than at the image level. Since the constructed datasets contain approximately balanced numbers of real and fake images, no class weighting, oversampling, undersampling, or synthetic sample generation was employed during training.

Furthermore, with each original image X, we compute a gradient map X_g using a transformation model (based on ResNet-50). All original images and their corresponding maps are passed into the feature extractor. The feature extractor consists of three separate components: RGB (Spatial), gradient, and frequency, each using two sets of convolutions to generate feature maps F_x, F_g, and F_r, respectively. We then combine them to generate a final feature map F_{map} of size (192, 224, 224). The final feature map F_{map} serves as input to the detection module. We empirically assess the performance of our framework across multiple systematic settings to investigate generalization.

3.1 Dataset Selection and Data Structuring

To ensure a fair and reliable evaluation of SGFF-Net, a comprehensive dataset construction pipeline was developed. One important criterion for dataset construction was to ensure that no identity information leaked. This means creating identity-disjoint datasets with no overlap of identities between datasets used for training, validation, testing, cross-model evaluation, cross-generator evaluation, real-world evaluation, and transformation model training.

Table 2 provides an overview of the three benchmark datasets used in this study, including their generation paradigm, source generative models, and the corresponding real-image datasets. These benchmark datasets are subsequently restructured into multiple identity-disjoint evaluation datasets to support the experimental settings described later in this section. The details of these datasets, presented in Tables 3 and 4, are described below.

The real images in the original benchmark datasets are derived from three sources: IMDB-WIKI [49], CelebA-HQ [50], and FFHQ [51]. However, only IMDB-WIKI and CelebA-HQ were used in this study because these two datasets provide identity annotations together with paired synthetic images. For each dataset, we constructed identity-level splits and then divided the datasets into multiple subsets to support different evaluation scenarios: (1) Intra-dataset, (2) Cross-model, (3) Cross-paradigm, (4) Multi-source, and (5) Real-world evaluation.

The DFF dataset was selected as the primary training dataset. Real image identity metadata were recovered from the original IMDB-WIKI dataset by matching image filenames with the official metadata files. The IMDB-WIKI dataset consists of two sub-datasets: IMDB and WIKI. After mapping all images, it was found that all

Table 2 Original benchmark datasets used to construct the evaluation datasets

DFF	DiffFace	DFFD
	Generation Paradigm	
Diffusion	Diffusion	GAN
	Generator Models	
SD v1.5 (T2I)	P2, DDPM, DDIM	PGGAN V1/V2
SD Inpainting	PNDM, LDM	StyleGAN
InsightFace	SD v1.5 (T2I)	StarGAN
	SD v1.5 (I2I)	FaceApp
	SD v2.1 (T2I)	
	SD v2.1 (I2I)	
	DiffSwap, Inpainting	
	Internet	
	Source of Real Images	
IMDb-WIKI	CelebA-HQ	CelebA-HQ, FFHQ

Table 3 Identity-aware partitioning of the real-image source datasets

Stage / Allocation	Images	IDs
	IMDb–WIKI	
Original	30,000	29,855
Preprocessed	25,605	25,522
DFF	22,455	22,375
Transformation Model	3,150	3,147
	CelebA-HQ	
Original	30,000	6,217
Preprocessed	24,388	5,799
DFFD_A	20,165	2,998
DFFD_B	1,596	1,127
Transformation Model	2,627	1,674

30,000 real DFF images originated from the WIKI subset, while none came from the IMDB dataset. After pre-processing, only 25,605 real images remained in the DFF dataset, of which only 25,522 had unique identities. Among these, 3,147 unique identities (3,150 images) were reserved for training the transformation model. The remaining 22,375 unique identities (22,455 images) were used to construct the identity-level train, validation, and test splits for DFF datasets.

For each real image, a single corresponding fake image was selected using a balanced sampling strategy. For DFF_C, one fake image was randomly chosen from either the SD v1.5 or the inpainting version, while the alternative version was used if the selected image was unavailable. All images belonging to the same identity were assigned exclusively to a single partition. Identity verification was then performed across all splits, confirming that there is no identity overlap between the training, validation, and testing sets.

Table 4 Structured datasets used across the evaluation settings

Dataset	Real	Fake*
	Intra-Dataset	
DFF_A	17,768	17,768
DFF_B	20,750	20,750
DFF_C	22,455	22,455
	Cross-Model	
DiffFace_A	3,000	3,000
DiffFace_B	3,000	3,000
DiffFace_C	3,000	3,000
DiffFace_D	3,000	3,000
DiffFace_E	3,000	3,000
DFFD_B	1,596	1,596
	Cross-Paradigm	
DFFD_A	20,165	21,729
	Transformation Model Training	
Transformation_Model	5,777	5,777
	Multi-Source	
MIX_ALL	42,620	44,184
	Real-World	
Real_World	200	200

*Fake-image sources: DFF_A: SD v1.5; DFF_B: SD Inpainting; DFF_C: SD v1.5/Inpainting; DiffFace_A: ADM, DDPM, DDIM, PNDM, LDM; DiffFace_B–E: ADM, DDPM, LDM, and PNDM, respectively; DFFD_A: PGGAN v1/v2; DFFD_B: StyleGAN. Transformation_Model combines reserved DFF and DFFD samples; MIX_ALL combines DFF_C and DFFD_A; Real_World contains independently collected web images.

The real image source of both DiffFace and DFFD is CelebA-HQ Since CelebA-HQ does not provide identity names, the official CelebAidentity annotations were obtained from the researchers who compiled this dataset and mapped to CelebA-HQ. The mapped dataset contained 30,000 images (6,217 unique identities). To eliminate identity leakage, all identities common between DFF and CelebA-HQ were removed. After filtering out duplicate identities, the final CelebA-HQ dataset contained 24,388 real images from 5,799 unique identities.,

For DiffFace, 3,000 real images were randomly selected from the filtered CelebA-HQ dataset. Their corresponding fake images were obtained from diffusion models, including Denoising Diffusion Probabilistic Model (DDPM) [52], Pseudo Numerical Methods for Diffusion Models (PNDM) [53], Latent Diffusion Models (LDM), Denoising Diffusion Implicit Models (DDIM) [41], and Ablated Diffusion Models (ADM) [54] to construct five evaluation datasets. Accordingly, DiffFace was partitioned into five subsets (A-E), each representing a different diffusion generator for cross-model evaluation.

The filtered CelebA-HQ identities were partitioned into three mutually exclusive groups. First is DFFD_A, which is used for model training and contains real images alongside synthetic images produced by PGGAN V1 and PGGAN V2. This dataset is partitioned at the identity level into training, validation, and testing subsets. The second group, called DFFD_B, contains real images paired with fakes generated by

StyleGAN and is reserved exclusively for cross-model evaluation (after training is carried out using DFFD_A). The third group, consisting of the remaining identities, is reserved exclusively for training the gradient transformation model.

The transformation model dataset was constructed by combining reserved images from the DFF and DFFD datasets. From DFF, these were the 3,150 real images reserved, along with their corresponding fake images from SD v1.5 and inpainting. From DFFD, these were the reserved 2,627 real images, and their corresponding fake images from ADM, DDIM, DDPM, LDM, PGGAN v1, PGGAN v2, and StyleGAN. The resulting merged dataset, consisting of 5,777 real images and 5,777 fake images, was used exclusively for training the gradient transformation model.

To evaluate the proposed framework on a more diverse dataset, a combined dataset called MIX_ALL was created by merging the splits of DFF_C and DFFD_A. Specifically, the training, validation, and test sets were formed by combining the respective partitions from both datasets. Since DFF_C and DFFD_A were constructed to contain mutually exclusive identities, the resulting MIX_ALL dataset also maintained zero identity overlap across all partitions.

Algorithm 1 Construction of the Real_World Evaluation Dataset

Require: Public identity sources, training identities $\mathcal{I}_{train}$, real/fake query templates
Ensure: Identity-disjoint Real_World dataset $\mathcal{D}_{RW}$

Collect and deduplicate public identities to obtain candidate set $\mathcal{I}$
Remove identities overlapping with $\mathcal{I}_{train}$
Randomly sample identities from the remaining candidate pool
for each identity $i \in \mathcal{I}$ **do**
    Generate real-image and AI-generated-image search queries
    Retrieve candidate images and associated metadata
    Filter candidates using source-domain and synthetic-content criteria
    Retain images containing one valid face using MTCNN
    Remove duplicate images using perceptual hashing
end for
Retain identities having valid candidates from both classes
Independently annotate retained images for identity, source reliability, image quality, and class validity
Resolve annotations using majority voting
Construct a balanced identity-disjoint dataset from the verified samples
return $\mathcal{D}_{RW}$

Finally, to assess SGFF-Net's robustness under real-world conditions, we constructed the Real_World dataset comprising real and AI-generated face images collected from public web search results. The dataset was constructed to ensure that its identities do not overlap with the datasets described above. The identity-disjoint retrieval and verification procedure for this dataset is summarized in Algorithm 1. The process begins by collecting public identities across business, sports, entertainment, science, and media. After cleaning and deduplication, 2,311 unique identities

are obtained. To prevent identity leakage, identities present in the training source data (IMDB-WIKI and CelebA-HQ) were removed, leaving approximately 1,198 unique unseen identities. From these, 398 identities were selected. Multiple search queries were generated for each identity by applying predefined templates for both Real and Fake image types (e.g., "[Name]" and "[Name] portrait" for real images, and "[Name] deepfake" and "[Name] AI generated" for AI-generated images).

Candidate images were retrieved automatically using Serper.dev, a Google Search API, and filtered using source-domain rules for real images and keyword/domain-based relevance criteria for AI-generated images. To improve source reliability during automated retrieval, real-image candidates were restricted to a predefined allow-list of sources, including Reuters, Getty, Wikimedia Commons, BBC, CNN, and government or educational domains, while user-upload platforms such as Pinterest, Imgur, and Reddit were excluded. AI-generated image candidates were retained only when their keyword/domain-based synthetic-content score exceeded the predefined threshold, while sources included in the real-image allow-list were excluded from AI-generated candidate selection.

All candidate images then underwent automated face verification using MTCNN, requiring a single detectable face satisfying the predefined minimum-size criterion, followed by perceptual-hash-based deduplication. After automated collection and filtering, 374 identities (1,891 images) remained with both valid real and AI-generated face-image candidates.

The retained candidates were independently examined by six annotators using a predefined annotation protocol covering identity match, source reliability, visible synthesis artifacts, and image quality. Each annotator independently rated all 1,891 candidate images as either "keep" or "discard". The final annotation for each image was determined by majority voting. A total of 1,166 images were labeled "keep", 523 were labeled "discard", and 202 were excluded because of a tie. From the retained images (283 identities), 200 identities were randomly selected, with one real and one fake image chosen at random for each selected identity when multiple candidates were available. This process yielded a curated dataset of 200 identities and 400 images (200 real and 200 AI-generated deepfakes). The complete metadata and code for the creation of this identity-disjoint and manually verified Real_World dataset are provided in a GitHub repository.

3.2 Preprocessing

The preprocessing module extracts and localizes face regions using MTCNN and resizes images to 224×224 pixels. The resulting images are stored in RGB JPEG (Joint Photographic Experts Group) format. Grayscale images are removed, and to eliminate visually similar images, duplicate detection was performed using Perceptual Hash (pHash). A 64-bit perceptual hash was generated for each image, and image pairs with a Hamming distance of ≤ 5 were identified as duplicates and removed. The dataset was then split at the identity level rather than the image level. Unique identities were split into 80% for training, 5% for validation, and 15% for testing. All images belonging to the same identity were assigned to only one split. After the split was

generated, identity and duplicate verification scripts were used to verify zero identity overlap between all dataset partitions (i.e., training, validation, and testing).

Another key part of the pre-processing module is that the original RGB image, X, is also processed by a transformation model to produce a gradient map X_g. The detailed description and the mathematical function for creating gradient maps are discussed below.

Prior studies have indicated that CNN gradient responses capture generator-induced artifacts that transfer more effectively across synthetic image generation methods than raw pixel representations [22, 23]. Following this observation, a ResNet-50-based transformation model is trained from scratch without pre-trained weights using a dedicated transformation dataset created from reserved images sampled from the DFF and DFFD datasets. These images are completely disjoint from those used for detector training, validation, and evaluation, ensuring that no images used during evaluation are used to train the transformation model.

Mathematically, a dataset is defined $X = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{3\times H\times W}$ represents an RGB image, $y_i \in \{0, 1\}$ denotes the class label (`Real` or `Fake`), N is the total number of samples, H represents the height of the image, and W denotes the width of the image. Also, given that the trained transformation model is denoted as $G(\cdot)$, each input image, x_i, is passed through the trained transformation model to obtain the corresponding output representation:

$$z_i = G(x_i) \tag{1}$$

where z_i is the output vector of the transformation model $G(.)$. Then, the output logits are summed, and the gradient of the sum (z_i) with respect to the input image x_i is computed. Finally, the gradient map of the summed output with respect to the input image is computed as:

$$M_i = \frac{\partial \sum z_i}{\partial x_i} \tag{2}$$

After training, the best-performing ResNet-50 checkpoint was frozen and used exclusively for gradient-map generation throughout all experiments. The gradients were calculated with respect to the original input image rather than intermediate feature representations by taking the derivative of the summed output logits. RGB inputs were converted into tensors within the range $[0, 1]$ without applying ImageNet-based mean and standard deviation normalization. Labels were used only during the initial training of the ResNet-50 transformation model, not during gradient map generation.

3.2.1 Selection of Frequency Representation

The choice of frequency representation method for frequency-enhanced images within the DWT block is based on an empirical evaluation of different frequency representations (i.e., FFT, DCT, DWT) to determine which performs best at detecting DM-generated deepfake images. Two categories of datasets are created: the original RGB and its grayscale version. Then, for each version, the image is converted to the frequency domain using 2D FFT, DCT, and DWT. A high-pass circular mask is

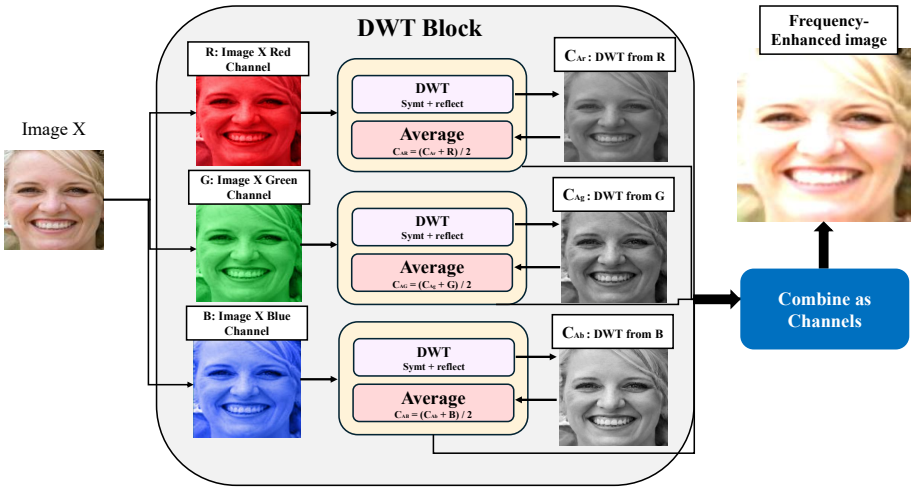


Fig. 2 DWT-based frequency extraction module, where each RGB channel is transformed independently and fused with its corresponding original channel to generate a frequency-enhanced RGB representation. Each RGB channel is decomposed using DWT, reconstructed into a frequency-enhanced representation, and then passed to the frequency feature extractor branch.

applied and reconstructed in the case of FFT and DCT. For DWT, the HH (High-High) components are extracted and resized to the original resolution. The selection of high-frequency components is motivated by the results of existing literature [25–28], which demonstrate that synthesis and reconstruction artifacts are more prominent in high-frequency regions. To determine the most effective representation, comparative experiments are conducted, and the results consistently show that RGB DWT-based representations achieve superior performance across most evaluation metrics, particularly in detecting diffusion-based manipulations (see Table A.1). Based on these findings, DWT is selected as the primary frequency transformation method in the DWT-block. The architecture of this module is illustrated in Fig. 2.

3.2.2 DWT Transformation

Initially, each input image is represented in RGB format. Instead of feeding this image directly into a convolutional network, the frequency branch first decomposes it into multiple frequency bands. A single-level two-dimensional DWT is applied to each channel. The DWT module decomposes the input signal into four distinct subbands: LL (low-frequency in both directions), LH (low-frequency horizontal, high-frequency vertical), HL(high-frequency horizontal, low-frequency vertical), and HH (high-frequency in both directions). This process is performed independently for each channel (red, green, and blue) to preserve each channel's unique frequency characteristics. The LL subbands represent a low-frequency approximation of the image and capture its global structure, such as smooth regions and overall shape. The LH, HL,

and HH subbands capture high-frequency details corresponding to horizontal and vertical edges, as well as fine-grained textures or noise, respectively. To evaluate the contribution of each component, multiple configurations are tested, including individual subbands (LL, HH), partial combinations (LL+HH), and full fusion of all subbands (LL+LH+HL+HH). Based on the comparative results (see Figure A.1), the fusion of all subbands (LL+LH+HL+HH configuration) was selected, as it had the highest accuracy and F1 score.

Additionally, the choice between different DWT wavelet families depends on the specific characteristics of the deepfake images and the requirements of the detection task. Therefore, the Haar, Symlet, and Biorthogonal wavelet types are assessed for deepfake detection. The best results are obtained using the Symlet wavelet type (see Table A.2). When applying DWT, borders need to be handled; for this purpose, different signal extension modes are used. The most common modes are zero padding, symmetric, and anti-reflect. All of these modes are tested (see Table A.2), and because the highest score is achieved with the reflective mode, it is implemented in the final design.

For each RGB channel, the DWT-transformed representation is averaged with its corresponding original channel using equal weights to generate a frequency-enhanced representation. This fusion preserves the spatial information of the original image while incorporating complementary frequency-domain characteristics extracted by the DWT. The resulting combined subband is reconstructed channel-wise, and the processed channels are merged to produce a three-channel frequency-enhanced image (see Fig. 2). This frequency-transformed image is subsequently converted into a tensor representation and passed through a dedicated convolutional branch to give $F_r \rightarrow (64, 224, 224)$. Finally, using channel-wise fusion, F‘ and F_r are combined to form a final feature map F_{map} of size $(192, 224, 224)$, which is then fed into the detection module.

3.3 Detection Network

Inspired by the ResNet [55] foundation model and the model developed in [25], a Dual Residual Network (DRN) architecture is selected for the detection module. Specifically, it is designed to efficiently learn discriminative representations, combining residual learning, dual-path feature learning, and Instance-Batch Normalization (IBN). The network has relatively few parameters and operates on fused multi-domain features, where the local path facilitates inter-channel feature interactions across the color, gradient, and frequency domains, while the global path captures spatial correlations in forgery artifacts. Additionally, it is lightweight because it incorporates IBN into the early layers, thereby reducing network depth and simplifying the classification head. These modifications significantly reduce computational complexity while improving generalization across diffusion and GAN-based deepfake datasets. Figure 3 visually represents the design of the detector module.

The DRN begins with a convolutional layer that transforms the initial 192-feature map into 64 channels using a 7×7 kernel with a stride of 2. This operation reduces spatial information in the fused features while capturing global contextual information. The output from this layer is then passed through an IBN layer followed by a

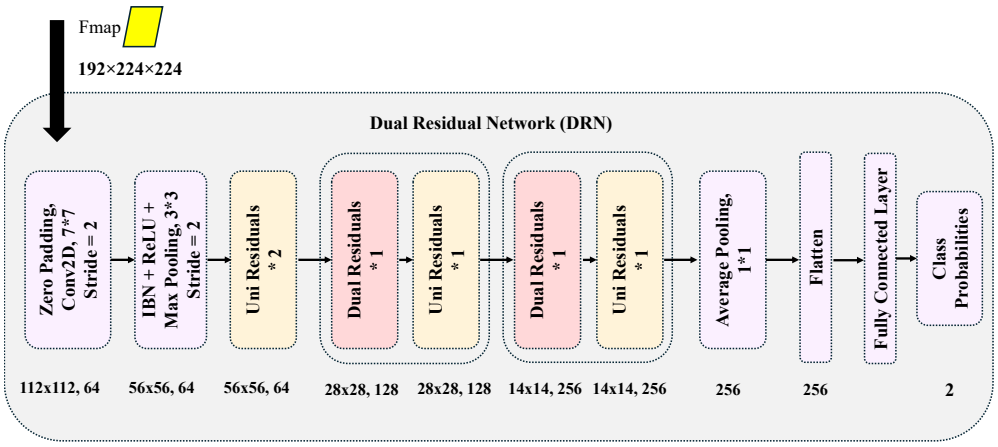


Fig. 3 Architecture of the Dual Residual Network (DRN) used for classification. The network combines local and global residual paths with IBN normalization to learn robust forgery representations from fused multi-domain features.

ReLU activation function. The IBN layer divides the feature channels into two groups: one half is normalized using Instance Normalization, and the other half is normalized using Batch Normalization. Following the initial encoding, the network applies a sequence of uni-residual blocks. Each block consists of a convolutional layer followed by normalization and ReLU activation. To improve robustness to face variation, the IBN is used in the initial layer of the uni-residual block.

Dual-residual paths are also added to the model to capture fine-grained and contextual information. The initial path is a local path that captures fine-grained details and artifact patterns using 1×1 convolutions, and the global path captures broader contextual information using stacked 3×3 convolutions. The outputs of these two paths are combined along with a residual connection from the input, and the result is passed through a ReLU activation. Global feature aggregation is performed by passing the resulting feature map through a global average pooling layer, producing a compact 256-dimensional feature vector. Finally, the pooled feature vector is fed into a fully connected layer that maps it to a 2-dimensional output corresponding to the binary class. A softmax function is applied to the output logits for the final prediction, followed by selecting the class with the highest probability.

3.4 Augmentation Techniques

Recent comparative analyses demonstrate that augmentation strategies can enhance cross-dataset generalization performance in a deepfake detection system [39, 46]. In line with this, multiple data augmentation techniques (such as JPEG compression, random resizing, cropping, rotations, horizontal flipping, color jitter, and Gaussian

Table 5 Online data augmentation settings used during training.

Augmentation	Configuration
JPEG Compression	Quality factor: 40–100
Random Resized Crop	224×224; scale: 0.8–1.0; ratio: 3/4–4/3
Random Horizontal Flip	Probability: 0.5
Random Rotation	$\pm 10^{\circ}$
Color Jitter	Brightness, contrast, saturation: 0.2
Gaussian Blur	Kernel size: 3; probability: 0.2

blur) are applied during training. Overall, the primary goal of augmentation is to increase data variety by applying different techniques to simulate the variation and diversity of real-world images.

To evaluate the impact of data augmentation on model generalization, two training configurations are investigated: one using the preprocessed facial images without augmentation and another using online data augmentation during training. Augmentation was applied only to the pre-cropped facial images during training to increase the diversity of training samples without explicitly expanding the stored dataset. Validation and test images were not augmented to ensure a consistent and unbiased evaluation. Random transformations generated different image realizations across training iterations. The augmentation pipeline was applied sequentially using the settings presented in Table 5.

4 Experimental Details

The following experiments are conducted to evaluate the performance and generalization capability of the proposed framework.

1. Intra-dataset evaluation
2. Cross-model evaluation (with and without augmentation)
3. Cross-paradigm evaluation (with and without augmentation)
4. Multi-source (in-distribution) evaluation (with and without augmentation)
5. Real-world evaluation

The framework is implemented using PyTorch and trained on a system equipped with an NVIDIA GeForce RTX 3060 (12 GB) GPU. The software environment included Python 3.9.25 and CUDA 12.8 for GPU acceleration. All source code is publicly available on GitHub for the benefit of the research community.

During training, the model is optimized with the AdamW optimizer, starting with an initial learning rate of 0.001, which is reduced to 0.0001. A batch size of 32 and 30 training epochs are used together with a weight decay factor of 0.0001. A ReduceLROnPlateau scheduler monitors the validation loss and reduces the learning rate by a factor of 0.5 after three consecutive epochs without improvement. Cross-Entropy loss with label smoothing (0.05) and a dropout rate of 0.25 are used to improve generalization.

Table 6 Intra-dataset evaluation of SGFF-Net on four benchmark datasets.

Train Set	Test Set	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓
DFF_A	DFF_A	99.02 ± 0.16	98.75 ± 0.25	99.29 ± 0.14	99.02 ± 0.16	99.93 ± 0.03	1.03 ± 0.17
DFF_C	DFF_C	95.55 ± 0.53	94.94 ± 0.76	96.24 ± 0.84	95.58 ± 0.53	99.17 ± 0.14	4.39 ± 0.61
DFFD_A	DFFD_A	99.91 ± 0.03	99.81 ± 0.08	99.96 ± 0.02	99.88 ± 0.04	100.00 ± 0.00	0.11 ± 0.05
MIX_ALL	MIX_ALL	97.80 ± 0.27	96.95 ± 0.63	98.22 ± 0.51	97.58 ± 0.29	99.73 ± 0.05	2.13 ± 0.26

For each independent run, the model checkpoint corresponding to the lowest validation loss was selected for final evaluation. Unless otherwise stated, all reported results correspond to the mean performance obtained over five independent training runs. Each run was initialized with a different random seed (14, 20, 24, 25, and 29). All runs for an experiment used identical datasets, hyperparameters, and evaluation settings; only the random seed was varied across runs. For each experimental setting, the mean and standard deviation are reported to assess the statistical reliability of the proposed framework.

Inference time was evaluated using a batch size of 32. Before recording time, 10 warm-up batches were executed to eliminate CUDA initialization overhead. CUDA synchronization was applied before and after the timed evaluation. The reported inference time represents the average per-image latency, obtained by dividing the total inference time by the total number of test images.

The performance metrics include accuracy, binary-averaged precision, recall, F1-score, ROC-AUC (Area Under the Receiver Operating Characteristic Curve), and Equal Error Rate (EER). Accuracy, precision, recall, and F1-score are computed from the binary confusion matrix, while ROC-AUC and EER are computed from the predicted confidence scores.

5 Results

The proposed Spatial-Gradient-Frequency Fusion Network (SGFF-Net) framework is evaluated under progressively more challenging settings to assess both detection performance and generalization capability. Unless otherwise stated, all experimental results are reported as mean ± standard deviation over five independent runs using different random seeds. Performance metrics are reported as percentages (%) except inference time, which is reported in milliseconds (ms). Higher values indicate better performance for Accuracy, Precision, Recall, F1-score, and ROC-AUC, whereas lower values indicate better performance for EER and inference time. First, intra-dataset experiments measure performance when training and testing are performed on the same data distribution. Next, cross-model and cross-paradigm evaluations assess robustness to unseen generators and generation paradigms. Multi-source experiments investigate the impact of training data diversity, while real-world evaluation measures generalization to unseen samples collected outside the benchmark datasets. Together, these experiments provide a comprehensive analysis of the effectiveness and robustness of SGFF-Net.

Table 7 Cross-model evaluation of SGFF-Net within the same generative paradigm.

Train Set	Test Set	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓
			Without Augmentation				
DFF_C	DiffFace*	54.55 ± 1.93	53.11 ± 1.49	79.97 ± 3.29	63.77 ± 0.69	59.50 ± 2.14	43.45 ± 1.66
DFFD_A	DFFD_B	99.98 ± 0.02	100.00 ± 0.00	99.96 ± 0.03	99.98 ± 0.02	100.00 ± 0.00	0.00 ± 0.00
			With Augmentation				
DFF_C	DiffFace*	62.07 ± 1.55	59.50 ± 1.51	78.81 ± 4.40	67.63 ± 1.77	68.57 ± 1.63	36.62 ± 1.32
DFFD_A	DFFD_B	99.76 ± 0.10	100.00 ± 0.00	99.52 ± 0.21	99.76 ± 0.10	100.00 ± 0.00	0.02 ± 0.03

*Mean A–E

5.1 Intra-dataset Performance Evaluation

Table 6 presents the results for training and testing on the same dataset. The results indicate that SGFF-Net achieves consistently high performance across both diffusion-based and GAN-based datasets, with accuracy exceeding 95.55% in all cases.

Among all GAN-based datasets, DFFD_A achieves the highest accuracy of 99.91% and a robust F1-score of 99.88%. Among the diffusion-based datasets, DFF_A achieves the highest accuracy at 99.02%. This is followed by DFF_C (95.55%), indicating that the SGFF-Net has effectively captured the underlying discriminative artifacts present in diffusion-based generators. In the case of the MIX_ALL dataset, which combines GAN- and diffusion-based generators, the accuracy is slightly lower, at 97.80%. This can be attributed to combining diverse generation techniques, which makes the classification task more challenging by adding complexity and variability to the data.

5.2 Cross-Model Evaluation

The results in Table 7 indicate that when SGFF-Net is trained on diffusion-based datasets and evaluated on the average performance across the DiffFace subsets (A–E), it exhibits a substantial drop in accuracy compared to the intra-dataset accuracy of 95.55%. Likewise, precision and ROC-AUC also decline significantly, indicating reduced discriminative capability. This behavior can be attributed to different diffusion models, such as unconditional ADM, DDPM, LDM, and PNDM, which introduce distinct artifact distributions different from those introduced by conditional models such as SD v1.5 text-to-image and SD Inpainting. These differences limit the transferability of features learned from a single source.

In contrast, SGFF-Net demonstrates excellent cross-model generalization between GAN-based generators, achieving 99.98% accuracy and F1-score when trained on DFFD_A (PGGAN) and tested on the DFFD_B (StyleGAN) dataset. It also achieves the highest ROC-AUC (100%) and the lowest EER (0%), demonstrating excellent transferability of SGFF-Net across these GAN-based generators.

The results also demonstrate the positive impact of data augmentation on diffusion-based cross-model evaluation. Specifically, augmentation improves accuracy from 54.55% to 62.07% (+7.52%), precision from 53.11% to 59.50% (+6.39%), F1-score from 63.77% to 67.63% (+3.86%), ROC-AUC from 59.50% to 68.57% (+9.07%), and reduces EER from 43.45% to 36.62%, corresponding to a 6.83% reduction.

Table 8 Cross-paradigm evaluation from diffusion-based to GAN-based datasets.

Train Set	Test Set	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓
			Without Augmentation				
DFF_C	DFFD_A	65.74 ± 1.03	56.34 ± 0.86	81.42 ± 3.03	66.56 ± 0.59	74.19 ± 1.79	30.41 ± 0.24
DFF_C	DFFD_B	69.53 ± 2.30	66.81 ± 3.42	78.52 ± 3.79	72.06 ± 1.04	76.89 ± 2.23	29.93 ± 2.15
			With Augmentation				
DFF_C	DFFD_A	69.01 ± 2.50	59.02 ± 2.85	80.41 ± 4.27	67.98 ± 2.17	77.34 ± 1.19	29.30 ± 1.03
DFF_C	DFFD_B	71.59 ± 1.01	70.16 ± 2.73	75.69 ± 4.86	72.67 ± 1.24	78.82 ± 1.26	28.05 ± 1.17

5.3 Cross-Paradigm Evaluation

To evaluate the ability of SGFF-Net to generalize across fundamentally different generation paradigms, two complementary experiments were conducted:

1. training on diffusion-based datasets and testing on GAN-based datasets, and
2. training on GAN-based datasets and testing on diffusion-based datasets.

The first experiment evaluates the transferability of representations learned from diffusion-based generators to GAN-generated images. Specifically, SGFF-Net is trained on the diffusion-based dataset (DFF_C) and evaluated on the GAN-based datasets (DFFD_A and DFFD_B). Table 8 shows that the framework achieves accuracies of 65.74% and 69.53% on DFFD_A and DFFD_B, respectively. Although these results represent a decrease compared with the corresponding intra-dataset performance, they demonstrate substantially stronger generalization than the cross-model setting, where training on DFF_C and testing on the unseen DiffFace dataset yields an accuracy of 54.55% (Table 7). For DFFD_A, the relatively high recall of 81.42% compared with the precision of 56.34% indicates that the model detects a large proportion of fake samples but also produces a high number of false positives. On DFFD_B, the precision and recall are more balanced, at 66.81% and 78.52%, respectively.

Data augmentation also improves cross-paradigm generalization. The accuracy increases from 65.74% to 69.01% on DFFD_A and from 69.53% to 71.59% on DFFD_B. ROC-AUC also increases from 74.19% to 77.34% and from 76.89% to 78.82%, respectively, while EER decreases from 30.41% to 29.30% on DFFD_A and from 29.93% to 28.05% on DFFD_B.

The second experiment evaluates reverse transfer, in which SGFF-Net is trained on the GAN-based dataset (DFFD_A) and evaluated on the diffusion-based datasets (DFF_A, DFF_B, and DFF_C). The results are presented in Table 9, with the SGFF-Net achieving accuracies of 76.05%, 66.98%, and 71.99% on DFF_A, DFF_B, and DFF_C, respectively. After augmentation, these accuracies increase to 82.32%, 70.68%, and 77.34%, respectively.

In comparison with the diffusion-to-GAN cross-paradigm evaluation, the GAN-to-diffusion setting demonstrates stronger cross-paradigm generalization. Overall, data augmentation consistently improves accuracy across the evaluated GAN-to-diffusion settings, with the highest accuracy of 82.32% achieved on DFF_A. These results indicate that incorporating data augmentation with multi-domain learning can further enhance the model's performance and its ability to generalize across different data distributions.

Table 9 Cross-paradigm evaluation from GAN-based to diffusion-based datasets.

Train Set	Test Set	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓
			Without Augmentation				
DFFD_A	DFF_A	76.05 ± 4.48	83.17 ± 2.91	65.43 ± 10.47	72.83 ± 6.90	85.49 ± 3.33	22.20 ± 3.22
DFFD_A	DFF_B	66.98 ± 2.79	67.35 ± 1.55	66.10 ± 10.39	66.35 ± 5.44	73.81 ± 2.88	32.63 ± 2.30
DFFD_A	DFF_C	71.99 ± 3.06	74.86 ± 2.16	66.44 ± 9.58	70.05 ± 5.31	80.43 ± 2.53	27.31 ± 2.44
			With Augmentation				
DFFD_A	DFF_A	82.32 ± 6.72	77.17 ± 8.64	92.89 ± 0.18	84.17 ± 5.09	91.07 ± 4.40	16.32 ± 4.88
DFFD_A	DFF_B	70.68 ± 5.60	64.61 ± 5.19	92.58 ± 0.77	76.03 ± 3.34	81.41 ± 5.68	26.29 ± 5.29
DFFD_A	DFF_C	77.34 ± 5.96	70.87 ± 6.36	93.72 ± 0.02	80.63 ± 4.14	87.56 ± 4.72	20.09 ± 4.84

Table 10 Single-source to mixed-source evaluation of SGFF-Net using MIX_ALL.

Train Set	Test Set	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓
			Without Augmentation				
DFF_C	MIX_ALL	78.76 ± 1.66	71.08 ± 2.53	89.37 ± 1.81	79.14 ± 1.04	89.60 ± 0.79	19.20 ± 0.87
DFFD_A	MIX_ALL	86.43 ± 1.48	87.21 ± 1.79	82.00 ± 5.14	84.42 ± 2.22	95.11 ± 0.62	13.18 ± 1.10
			With Augmentation				
DFF_C	MIX_ALL	80.99 ± 1.15	74.10 ± 2.47	89.09 ± 2.38	80.85 ± 0.62	91.00 ± 0.43	17.40 ± 0.51
DFFD_A	MIX_ALL	85.74 ± 3.38	81.28 ± 4.15	89.25 ± 8.07	84.85 ± 3.87	94.06 ± 1.72	13.34 ± 3.49

5.4 Multi-source Evaluation

To assess the effectiveness of heterogeneous multi-source training for improving model performance capability, we conduct two complementary in-distribution evaluations:

1. training on a single dataset and testing on a mixed-source dataset (MIX_ALL), and
2. training on the mixed-source (MIX_ALL) dataset and evaluating on its multiple constituent source datasets.

These experiments assess the effect of heterogeneous source diversity rather than out-of-distribution (OOD) generalization. MIX_ALL contains source distributions represented in its constituent datasets; therefore, the mixed-to-single-source setting is an example of in-distribution evaluation.

The results from the first experiment are presented in Table 10. Training on DFF_C and evaluating on MIX_ALL produces an accuracy of 78.76%, with ROC-AUC of 89.60% and EER of 19.20%. With augmentation, accuracy increases to 80.99% and ROC-AUC to 91.00%, indicating improved discrimination in the mixed-source setting. When DFFD_A is used as the sole training source, the results are higher than those obtained with DFF_C. The model achieves 86.43% accuracy, 84.42% F1-score, and 95.11% ROC-AUC on MIX_ALL without augmentation. With augmentation, recall increases from 82.00% to 89.25%, while precision decreases from 87.21% to 81.28%. Thus, data augmentation primarily benefits the diffusion-based model (DFF_C), improving accuracy by 2.23% and reducing the EER by 1.8%. In contrast, the GAN-based model (DFFD_A) exhibits only marginal changes and does not provide a uniform improvement across all metrics in this setting.

In the reverse evaluation, SGFF-Net is trained on MIX_ALL and evaluated on its constituent individual datasets (including DFF_A, DFF_B, DFF_C, DFFD_A, and DFFD_B). The performance of SGFF-Net is substantially better, as presented in Table 11 because of the in-distribution evaluation. The model achieves 98.87%, 97.65%, and 95.52% accuracy on DFF_A, DFF_B, and DFF_C, respectively. The corresponding ROC-AUC values are 99.90%, 99.75%, and 99.12%, with EER values below 4.4%.

Table 11 Mixed-source to single-source evaluation of SGFF-Net using MIX_ALL

Train Set	Test Set	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓
MIX_ALL	DFF_A	98.87 ± 0.28	98.54 ± 0.66	99.20 ± 0.28	98.87 ± 0.27	99.90 ± 0.03	1.07 ± 0.17
MIX_ALL	DFF_B	97.65 ± 0.34	95.99 ± 0.75	99.45 ± 0.20	97.69 ± 0.32	99.75 ± 0.09	1.78 ± 0.29
MIX_ALL	DFF_C	95.52 ± 0.54	94.43 ± 1.12	96.76 ± 0.96	95.57 ± 0.52	99.12 ± 0.15	4.37 ± 0.61
MIX_ALL	DFFD_A	99.94 ± 0.02	99.95 ± 0.03	99.92 ± 0.04	99.93 ± 0.02	100.00 ± 0.00	0.03 ± 0.03
MIX_ALL	DFFD_B	99.89 ± 0.04	99.97 ± 0.05	99.80 ± 0.06	99.89 ± 0.04	100.00 ± 0.00	0.05 ± 0.03

Table 12 Real-world evaluation of SGFF-Net under different training-source configurations.

Train Set	Test Set	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓
			Without Augmentation				
DFFD_A	Real_World	49.80 ± 2.77	49.79 ± 1.70	82.10 ± 8.17	61.94 ± 3.63	51.18 ± 4.54	49.50 ± 2.72
DFF_C	Real_World	55.60 ± 2.50	54.43 ± 1.55	67.30 ± 5.77	60.05 ± 4.38	56.96 ± 3.11	46.50 ± 3.08
MIX_ALL	Real_World	58.65 ± 2.44	55.50 ± 1.79	88.50 ± 4.18	68.16 ± 1.31	62.47 ± 3.67	41.20 ± 2.56
			With Augmentation				
MIX_ALL	Real_World	63.15 ± 2.01	58.68 ± 1.75	89.50 ± 2.60	70.85 ± 0.85	70.42 ± 1.83	37.00 ± 1.84

Similarly, the model achieves 99.94% and 99.89% accuracy on DFFD_A and DFFD_B, respectively, with ROC-AUC values of 100.00% and EER values below 0.1%.

Overall, Tables 10 and 11 provide two main insights. First, exposure to diverse source distributions during training improves performance across the represented source distributions and enables SGFF-Net to learn more robust discriminative representations. Second, models trained on a single source remain more sensitive to the composition of the evaluation distribution.

5.5 Real-World Data Evaluation

The results of the Real_World evaluation in Table 12 indicate that SGFF-Net benefits from heterogeneous training sources, with performance improving from 49.80% accuracy when trained with GAN-based data to 58.65% when trained with mixed GAN- and diffusion-based data. When augmentation is applied along with heterogeneous training sources, it further improves performance to 63.15% accuracy and 70.42% ROC-AUC while reducing EER to 37.00%. Although the performance remains lower than in the intra-dataset and cross-paradigm settings, the 63.15% accuracy is still informative, indicating that real-world domain shifts remain a substantial challenge, even for a multi-domain fusion detector. The ROC-AUC improvement with augmentation (+7.95%) suggests that augmentation helps the model better separate real and AI-generated samples across decision thresholds. The reduction in EER from 49.50% under single-source GAN training to 37% when mixed-source training is combined with augmentation indicates that greater source diversity and training variability jointly improve transfer to real-world data. However, the remaining error rate shows that these strategies do not eliminate the real-world distribution gap.

To further gain a deeper intuition of SGFF-Net's behavior under real-world conditions, a qualitative examination of the misclassified samples from the Real_World dataset was conducted. Figure 4 illustrates representative examples of failure cases, i.e., false positives (real images predicted as fake) and false negatives (fake images predicted as real), along with the model's prediction confidence for each case. Among the examined false-positive cases, several images exhibit unusual lighting, low resolution, or strong contrast. In contrast, the false-negative examples generally contain



Fig. 4 Qualitative failure cases of SGFF-Net trained on MIX_ALL with augmentation and evaluated on the Real_World benchmark. Top row: real images misclassified as fake (false positives). Bottom row: AI-generated images misclassified as real (false negatives).

realistic facial textures with few visually apparent synthesis artifacts, including several predictions with confidence exceeding 0.85.

5.6 Inference Time Evaluation

Inference latency of SGFF-Net across all evaluation settings is presented in Table 13. The average per-image latency remains within a narrow and stable range of 7.13–7.68 ms across intra-dataset, cross-model, cross-paradigm, multi-source, and real-world evaluations. The lowest latency is observed for the multi-source setting with augmentation (7.13 ms), whereas the highest latency is observed for the real-world evaluation with augmentation (7.68 ms). This stability is expected because the SGFF-Net architecture, input configuration, and batch size remain fixed during inference. Changing the evaluation dataset alters the input distribution but does not change the operations executed by the network.

6 Discussion

The experimental results demonstrate that SGFF-Net achieves strong in-distribution performance and maintains varying degrees of transferability under cross-model, cross-paradigm, and real-world distribution shifts. These results also provide valuable insights into architectural design choices, the nature of domain shift across generative models, and the role of data diversity and augmentation in achieving generalization. This section synthesizes these insights and connects them to the design decisions made for the proposed framework.

Table 13 Inference latency of SGFF-Net across different evaluation settings.

Evaluation	Train Set	Test Set	Augmentation	Inference Time (ms)↓
Intra-Dataset	DFF_A	DFF_A	Without	7.38 ± 0.06
	DFF_C	DFF_C	Without	7.31 ± 0.09
	DFFD_A	DFFD_A	Without	7.31 ± 0.04
	MIX_ALL	MIX_ALL	Without	7.22 ± 0.05
Cross-Model	DFF_C	DiffFace*	Without	7.38 ± 0.09
	DFFD_A	DFFD_B	Without	7.57 ± 0.20
	DFF_C	DiffFace*	With	7.32 ± 0.03
	DFFD_A	DFFD_B	With	7.44 ± 0.11
Cross-Paradigm (Diffusion → GAN)	DFF_C	DFFD_A	Without	7.30 ± 0.01
	DFF_C	DFFD_B	Without	7.33 ± 0.27
	DFF_C	DFFD_A	With	7.27 ± 0.11
	DFF_C	DFFD_B	With	7.52 ± 0.28
Cross-Paradigm (GAN → Diffusion)	DFFD_A	DFF_A	Without	7.41 ± 0.20
	DFFD_A	DFF_B	Without	7.29 ± 0.05
	DFFD_A	DFF_C	Without	7.28 ± 0.03
	DFFD_A	DFF_A	With	7.33 ± 0.05
	DFFD_A	DFF_B	With	7.30 ± 0.05
	DFFD_A	DFF_C	With	7.25 ± 0.04
Multi-source (Single → Mixed)	DFF_C	MIX_ALL	Without	7.27 ± 0.20
	DFFD_A	MIX_ALL	Without	7.30 ± 0.24
	DFF_C	MIX_ALL	With	7.16 ± 0.52
	DFFD_A	MIX_ALL	With	7.13 ± 0.08
Multi-source (Mixed → Single)	MIX_ALL	DFF_A	Without	7.36 ± 0.06
	MIX_ALL	DFF_B	Without	7.41 ± 0.26
	MIX_ALL	DFF_C	Without	7.26 ± 0.06
	MIX_ALL	DFFD_A	Without	7.36 ± 0.21
	MIX_ALL	DFFD_B	Without	7.13 ± 0.48
Real-World	DFFD_A	Real_World	Without	7.48 ± 0.14
	DFF_C	Real_World	Without	7.62 ± 0.11
	MIX_ALL	Real_World	Without	7.62 ± 0.25
	MIX_ALL	Real_World	With	7.68 ± 0.22

*Mean A–E

6.1 Architectural Design Insights

The architectural design of SGFF-Net is based on a series of ablation studies. Table 14 presents the results of feature-fusion ablation, which directly evaluates the contribution of the proposed multi-domain representation. Further experiments regarding the selection of frequency representation, DWT wavelet and boundary mode, and DWT subbands are reported in Appendix A.

The work of [25] inspires this study. However, the proposed framework is developed incrementally with design choices based on empirical evaluation. First, different frequency representations (FFT, DCT, and DWT) and image representations (RGB and grayscale) are evaluated to identify the most suitable frequency representation for detecting diffusion-generated deepfakes. The detailed results presented in Table A.1 indicate that DWT-based representations outperform both FFT and DCT for grayscale and RGB images. Compared with FFT, DWT improves accuracy by 5.17% for grayscale images and 6.60% for RGB images while also reducing EER. This result suggests that the localized multi-resolution representation provided by DWT is more effective at capturing the discriminative artifacts present in the evaluated

Table 14 Ablation of feature-domain combinations using detection accuracy.

Train Set	Test Set	FREQ	GRD	RGB+GRD	GRD+FREQ	RGB+FREQ	RGB+GRD+FREQ
MIX_ALL	MIX_ALL	93.59	80.91	95.98	90.70	95.74	**97.80**
DFF_A	DFF_A	95.18	86.15	95.65	98.23	**99.57**	99.02
DFF_C	DFF_C	94.94	84.89	94.45	95.41	95.37	**95.55**
DFFD_A	DFFD_A	89.59	81.62	90.58	93.40	96.10	**99.91**

diffusion-generated images than the global frequency representations provided by FFT and DCT.

The same ablation study indicates that RGB inputs consistently outperform grayscale inputs, suggesting that color information provides additional forensic cues beyond those captured by luminance information alone. This finding supports the use of RGB representations in the final feature extraction module.

The effectiveness of the frequency branch also depends on the wavelet family and boundary handling strategy. As reported in Table A.2, the combination of the Symlet wavelet with Reflect boundary extension achieves the highest ROC-AUC among the evaluated configurations. The DWT subband analysis further indicates that combining all four subbands (LL, LH, HL, and HH) yields the strongest overall performance across the evaluated settings, as illustrated in Figure A.1.These findings support the use of Symlet with Reflect boundary handling and the full LL+LH+HL+HH subband configuration in SGFF-Net.

The feature-fusion ablation study directly investigates whether combining spatial, gradient, and frequency-domain representations provides an advantage over individual and pairwise feature representations.

Table 14 shows that the complete RGB+GRD+FREQ fusion generally provides strong performance and achieves the highest accuracy on MIX_ALL, DFF_C, and DFFD_A, while on DFF_A, RGB+FREQ performs slightly better.

The improvement is particularly evident on DFFD_A, where the complete fusion achieves an accuracy of 99.91%, compared to 89.59% with frequency features alone and 81.62% with gradient features alone. Moreover, the complete fusion outperforms every pairwise configuration across the evaluated datasets. These results indicate that the spatial, gradient, and frequency branches provide complementary rather than redundant forensic information, empirically supporting the three-branch feature representation utilized in SGFF-Net.

6.2 Comparison with Single-Domain Baselines

For a fair and controlled comparison, both the original DRN [25] and the state-of-the-art LGrad [22] architectures were reproduced and trained using the same experimental protocol and evaluation splits as the proposed SGFF-Net. The only differences between the methods being compared are their preprocessing strategies and network architectures. The original DRN employs offline FFT-based frequency preprocessing on grayscale images, and LGrad utilizes offline gradient images generated by a fixed, pre-trained transformation model before feeding them to a ResNet classifier. In contrast, SGFF-Net integrates spatial, gradient, and DWT-based frequency representations into a unified detection framework and separately trains a transformation model for gradient map generation. The comparison of baseline DRN, LGrad, and SGFF-Net for

Table 15 Controlled comparison of DRN, LGrad, and SGFF-Net across intra-dataset, cross-model, and cross-paradigm settings. All methods are trained on DFF_C without augmentation.

Evaluation	Metric	DRN	LGrad	SGFF-Net
Intra-Dataset DFF_C → DFF_C	Accuracy	82.78 ± 0.22	70.47 ± 0.72	**95.55 ± 0.53**
	Precision	84.02 ± 0.60	71.98 ± 2.05	**94.94 ± 0.76**
	Recall	80.96 ± 1.41	67.28 ± 4.22	**96.24 ± 0.84**
	F1-score	82.45 ± 0.44	69.46 ± 1.58	**95.58 ± 0.53**
	ROC-AUC	90.91 ± 0.12	77.47 ± 0.70	**99.17 ± 0.14**
	EER ↓	16.83 ± 0.07	29.23 ± 0.76	**4.39 ± 0.61**
Cross-Model DFF_C → DiffFace(A-E)	Accuracy	51.26 ± 0.43	48.22 ± 0.73	**54.55 ± 1.93**
	Precision	52.41 ± 0.73	46.60 ± 1.38	**53.11 ± 1.49**
	Recall	27.88 ± 6.95	25.26 ± 8.23	**79.97 ± 3.29**
	F1-score	36.03 ± 6.08	32.26 ± 7.08	**63.77 ± 0.69**
	ROC-AUC	52.87 ± 0.41	47.70 ± 0.64	**59.50 ± 2.14**
	EER ↓	47.51 ± 0.25	51.04 ± 0.42	**43.45 ± 1.66**
Cross-Paradigm DFF_C → DFFD_A	Accuracy	44.70 ± 2.17	48.65 ± 1.84	**65.74 ± 1.03**
	Precision	30.47 ± 3.41	32.35 ± 2.25	**56.34 ± 0.86**
	Recall	29.42 ± 7.22	25.77 ± 8.60	**81.42 ± 3.03**
	F1-score	29.80 ± 5.24	28.34 ± 6.43	**66.56 ± 0.59**
	ROC-AUC	39.93 ± 2.87	43.94 ± 1.07	**74.19 ± 1.79**
	EER ↓	57.37 ± 2.39	54.25 ± 0.96	**30.41 ± 0.24**

intra-dataset, cross-model, and cross-paradigm evaluation is presented in Table 15. The results indicate that SGFF-Net consistently achieves the highest performance in the intra-dataset evaluation, with an accuracy of 95.55%, an F1-score of 95.58%, an ROC-AUC of 99.17%, and the lowest EER of 4.39%, substantially outperforming both DRN and LGrad.

In the more challenging cross-model evaluation, all three models experience a notable decrease in performance, but SGFF-Net still achieves the highest Recall (79.97%), F1 (63.77%), and the lowest EER (43.45%). SGFF-Net's performance indicates that multi-domain representations retain more discriminative information when evaluated on unseen diffusion generators than single-domain DRN and LGrad representations.

For cross-paradigm evaluation, SGFF-Net consistently achieved the highest performance with 65.74% accuracy, 66.56% F1-score, and 74.19% ROC-AUC, outperforming both DRN and LGrad. The computational implications of this multi-domain design are discussed separately in Section 6.6.

6.3 Open-World Detector Comparison

To contextualize SGFF-Net against a recent open-world AI-generated image detector, we include an evaluation of the official implementation of UniversalFakeDetect (UFD)[34] using the pretrained weights released by the authors. UFD employs a frozen CLIP ViT-L/14 encoder pretrained on approximately 400 million image-text pairs, and it is evaluated here in its zero-shot setting. In contrast, SGFF-Net is trained on the corresponding training split using task-specific spatial, gradient, and DWT-based forensic representations. The performance comparison given in Table 16 should therefore be interpreted as a comparison between zero-shot foundation-model transfer and task-specific forensic training rather than as a controlled head-to-head comparison under identical training conditions.

Table 16 Comparison of zero-shot UFD and dataset-trained SGFF-Net across the evaluated datasets.

Method	Train Set	Test Set	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓	Inference (ms)↓
UFD	Pretrained*	DFFD_A	90.50	94.50	86.00	90.05	96.71	10.90	121.77
SGFF-Net	DFFD_A	DFFD_A	99.91	99.81	99.96	99.88	100.00	0.11	7.31
UFD	Pretrained*	DFF_C	70.50	79.28	55.50	65.29	81.17	26.50	111.04
SGFF-Net	DFF_C	DFF_C	95.55	94.94	96.24	95.58	99.17	4.39	7.31
UFD	Pretrained*	MIX_ALL	80.50	87.88	70.75	78.39	89.66	18.00	112.37
SGFF-Net	MIX_ALL	MIX_ALL	97.80	96.95	98.22	97.58	99.73	2.13	7.22

*Zero-shot

UFD demonstrates strong zero-shot capability, achieving ROC-AUC values of 96.71%, 81.17%, and 89.66% on DFFD_A, DFF_C, and MIX_ALL, respectively. Its comparatively lower performance on DFF_C (70.50% accuracy) than on DFFD_A (90.50% accuracy) further suggests that strong pretrained semantic representations do not eliminate sensitivity to the target generator distribution. SGFF-Net's ROC-AUC values are higher than 99%, but it also records its lowest accuracy of 95.55% for the DFF_C dataset.

6.4 Understanding Generalization Across Generators and Paradigms

A clear hierarchy emerges across all evaluation settings. Performance is highest in intra-dataset evaluation, followed by cross-paradigm, cross-model, and finally real-world evaluation. This progressive degradation demonstrates that the difficulty of deepfake detection increases as the gap between training and testing distributions widens.

Cross-model evaluation reveals that transferability is strongly generator-dependent. SGFF-Net transfers effectively from PGGAN to StyleGAN, whereas substantially lower performance is observed when a detector trained on DFF_C is evaluated on unseen diffusion generators in DiffFace. Augmentation improves the latter setting but does not close the performance gap.

Although SGFF-Net significantly enhances deepfake detection performance, it does not eliminate the generalization gap. In fact, feature richness alone appears insufficient to achieve robust generalization under all conditions. The results from Tables 6 to 12 indicate that performance remains partially dependent on the distribution of the training data, and SGFF-Net continues to struggle when confronted with generators exhibiting fundamentally different generation characteristics.

The controlled baseline comparison further shows that although cross-paradigm transfer remains substantially more challenging than in-distribution detection, SGFF-Net retains more discriminative capability than DRN and LGrad under paradigm shift. This indicates that combining complementary spatial, gradient, and frequency representations improves the transferability of forensic cues, although it does not make these representations generator-invariant.

The results further indicate asymmetric cross-paradigm transfer: a model trained on GAN-based data appears to transfer somewhat better to diffusion-based data than a model trained on diffusion data transfers to GAN-based data. This suggests that

the discriminative cues learned from the two paradigms are not equally transferable across the paradigm boundary. Consequently, cross-paradigm evaluation remains one of the most challenging controlled settings despite the use of multi-domain features.

The real-world evaluation provides further evidence that generalization remains an open problem. The substantial gap between benchmark and real-world performance indicates that the learned representations are not yet fully invariant to unseen data distributions. Therefore, improving generator-invariant feature learning remains an important research direction.

6.5 Impact of Data Augmentation and Multi-Source Training

The experimental results also demonstrate the importance of data diversity and training variability for generalization. Augmentation can improve transferability, but its effect depends on the source–target distribution shift and is not uniform across evaluation settings. Specifically, it helps by introducing variability such as noise, compression artifacts, illumination changes, and geometric transformations during training. This exposes SGFF-Net to more robust and invariant feature representations rather than memorizing dataset-specific patterns. The effect is particularly evident under challenging cross-model and cross-paradigm evaluations.

Therefore, the augmentation results revealed several notable trends. First, the effect of augmentation is distribution-dependent rather than uniformly beneficial across all evaluation settings. For the DFF_C-to-DiffFace cross-model evaluation (see Table 7), augmentation increases accuracy by 7.52% and ROC-AUC by 9.07 % while reducing EER by 6.83 %. It also substantially improves sensitivity and overall discrimination for GAN-to-diffusion transfer (see Table 9), particularly in terms of overall discrimination (ROC-AUC) and error rate (EER), albeit at the expense of a precision–recall trade-off. In contrast, in the single-to-mixed-source (see Table 10) and DFFD_A-to-DFFD_B cross-generator settings (see Table 7), the experimental data show source-dependent changes rather than a consistent improvement. Second, the largest benefits are observed under cross-model and cross-paradigm shifts, indicating that the magnitude of improvement depends on the source–target distribution gap and the available room for improvement.

The cross-paradigm evaluation further demonstrates the effect of augmentation under larger distribution shifts. When SGFF-Net is trained on GAN-based DFFD_A and evaluated on unseen diffusion-based datasets (see Table 9, augmentation improves accuracy across all three target datasets. The magnitude of improvement varies across target distributions, but the consistent gains in accuracy and reductions in EER indicate that augmentation improves transfer from GAN-based to unseen diffusion-based generators.

The results also demonstrate the importance of training-data diversity. Training on MIX_ALL exposes the detector to both GAN- and diffusion-based source distributions and produces strong performance on its constituent datasets (see Tables 10 and 11). However, because these evaluations are in-distribution, they should not be interpreted as evidence of generalization to unseen generators. The real-world results provide a stronger test of whether heterogeneous source exposure improves transfer beyond the represented benchmark distributions.

Table 17 Computational complexity of DRN, LGrad, and SGFF-Net.

Method	Params. (M)	Train. (M)	MACs (G)	FLOPs (G)
DRN	4.84	4.84	1.12	2.23
LGrad	23.51	23.51	4.13	8.26
SGFF-Net	26.18	2.66	11.30	22.60

During real-world evaluation, when SGFF-Net is trained on the MIX_ALL dataset, augmentation improves accuracy from 58.65% to 63.15%, increases ROC-AUC from 62.47% to 70.42%, and reduces EER from 41.20% to 37.00% (see Table 12). However, this performance is considerably lower than that observed on the benchmark distributions. This indicates that heterogeneous training and augmentation improve transferability to real-world samples without eliminating the domain gap introduced by unseen data sources and generation conditions.

Prior studies have also investigated augmentation for improving deepfake detection robustness [39, 45]. However, their experimental settings differ from those considered here. Ameer et al. evaluated conventional augmentation using the DFDC dataset [40], while Chen et al. investigated diffusion-based augmentation and reported an AUC improvement from 73.12% to 78.02% on Celeb-DF [45, 56]. In contrast, this work evaluates augmentation under cross-model, cross-paradigm, multi-source, and real-world settings. Direct numerical comparison is not appropriate due to differences in datasets, detection models, and augmentation strategies; however, the findings provide complementary evidence that augmentation can improve robustness under distribution shift. The results of this study further indicate that the magnitude of this improvement depends on the source–target distribution shift.

6.6 Computational Efficiency and Practical Deployment

The computational evaluation considers both architectural complexity and empirical inference latency. Table 13 reports execution time under the standardized hardware and timing setup. The inference time remains relatively stable across the evaluated settings, with an average per-image latency of approximately 7–8 ms. This stability is expected because the SGFF-Net architecture, input configuration, and batch size remain fixed during inference; changing the evaluation dataset alters the input distribution but does not change the operations executed by the network. The observed latency provides an initial indication that the additional multi-domain processing remains computationally manageable on the evaluated GPU, although deployment suitability will depend on the target hardware, workload, and operational constraints.

The computational trade-off associated with multi-domain feature fusion is summarized in Table 17. This table also reports the model parameters, MACs, and FLOPs. DRN is the lightest model, with 4.84 million parameters and 2.23 GFLOPs, while LGrad requires 23.51 million parameters and 8.26 GFLOPs. SGFF-Net has the highest overall computational requirement, with 26.18 million parameters and 22.60 GFLOPs. However, only 2.66 million parameters are trainable because the gradient transformation model remains frozen during detector training. SGFF-Net introduces additional

Table 18 Inference latency of DRN, LGrad, and SGFF-Net

Evaluation	Method	Inference (ms)↓
Intra-Dataset	DRN	1.10 ± 0.06
	LGrad	1.70 ± 0.01
	SGFF-Net	7.31 ± 0.09
Cross-Model	DRN	1.13 ± 0.12
	LGrad	1.71 ± 0.003
	SGFF-Net	7.38 ± 0.09
Cross-Paradigm	DRN	1.03 ± 0.08
	LGrad	1.70 ± 0.005
	SGFF-Net	7.30 ± 0.01

computational cost relative to DRN and LGrad because it jointly processes spatial, gradient, and frequency-domain representations.

A similar trade-off is reflected in the inference latency reported in Table 18. DRN has the lowest inference latency, followed by LGrad, while SGFF-Net requires approximately 7.3 ms per image. Thus, the improved cross-generator performance of SGFF-Net is obtained at the cost of higher inference complexity, representing a performance–efficiency trade-off rather than a computational advantage over the baselines.

6.7 Limitations and Future Directions

Several limitations of this study should be acknowledged. First, the classification backbone (DRN) is inherently designed for single-domain spatial feature learning and might not fully exploit the complementary relationships between RGB, gradient, and frequency-domain representations. Second, the feature fusion strategy adopted in the proposed framework relies on simple concatenation, which does not explicitly model inter-feature interactions or attention-based weighting. Third, the frequency branch relies on fixed DWT decomposition without adaptive selection of informative sub-bands. Our objective in this work was not to design an adaptive frequency decomposition method, but rather to systematically investigate which frequency representations provide the strongest performance within the evaluated deepfake detection settings.

The benchmark datasets used in this work draw their real images from different source distributions: DFF uses IMDb-WIKI, whereas DFFD and DiffFace use CelebA-based images. Consequently, the observed performance differences may reflect both generator-paradigm variation and shifts in the underlying real-image distribution. Performing an ideal analysis in which the same real images are combined with outputs from both GAN- and diffusion-based generators would require developing a newly curated benchmark dataset, which is outside the scope of this study. Our objective is to evaluate model generalization across established public datasets, where both generator and data-distribution shifts are naturally present. We therefore acknowledge this as a limitation and consider it an important avenue for future investigation.

These limitations suggest several promising directions for future work. More advanced backbones based on attention mechanisms and transformers may better exploit cross-domain interactions among spatial, gradient, and frequency representations and should be explored. Similarly, adaptive frequency analysis techniques that dynamically select informative frequency components rather than relying on predefined DWT sub-bands, domain adaptation methods, and self-supervised learning approaches should also be investigated to assess their potential to improve robustness to unseen generators. Finally, expanding the diversity of training data, constructing a new benchmark that uses the same real images to generate both GAN and diffusion models, and explicitly designing domain-invariant representations remain important directions for achieving stronger real-world generalization.

7 Conclusion

This study investigates generalizable deepfake image detection across both GAN-based and diffusion-based generators and proposes SGFF-Net, a multi-domain feature fusion framework that integrates spatial (RGB), gradient, and DWT-based frequency representations within a DRN-based detector. The design of SGFF-Net is validated by ablation studies that establish the effectiveness of frequency representation and configuration, and demonstrate the complementary contributions of the three feature domains.

SGFF-Net is evaluated under progressively varied distribution settings, including intra-dataset, cross-model, cross-paradigm, multi-source, and an independently collected real-world dataset. The results indicate that integrating complementary spatial, gradient, and frequency representations enhances the transferability of forensic cues compared to the single-domain DRN and LGrad baselines. However, this transferability is still dependent on the generator. In particular, the asymmetric behavior observed in transfers from GAN-to-diffusion and diffusion-to-GAN indicates that robustness within one generative paradigm does not necessarily translate to the other.

The experiments further indicate that generalization depends on both feature representation and the composition of the training data. Heterogeneous multi-source training improves performance across the represented source distributions and provides better transfer to real-world data, while the effect of augmentation varies with the source–target distribution shift and becomes more pronounced under challenging shifts. Nevertheless, the performance degradation observed on independently collected real-world images demonstrates that combining feature diversity, source diversity, and augmentation does not fully bridge the domain gap.

These findings suggest that achieving generalizable deepfake detection cannot be accomplished through feature fusion alone. To ensure robustness against evolving generative models, representations must capture complementary forensic evidence while minimizing reliance on generator- and source-specific distributions. Future work will investigate more explicit cross-domain feature interactions, adaptive frequency representations, and domain-invariant representation learning. Moreover, controlled benchmarks in which GAN- and diffusion-generated images are constructed from the

Table A.1 Ablation study on frequency-domain representations using the DFF_A dataset. The performance of Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), and Discrete Wavelet Transform (DWT) representations is evaluated using both grayscale and RGB inputs.

Input Type	Method	Acc.↑	Prec.↑	Rec.↑	F1↑	ROC-AUC↑	EER↓	Inference(ms)↓
Grayscale	FFT	83.82	86.26	80.47	83.26	91.85	15.83	0.80
	DCT	79.39	76.06	85.79	80.63	87.77	20.18	**0.58**
	DWT	**88.99**	**89.56**	**88.27**	**88.91**	**95.15**	**11.08**	2.40
	DWT vs. FFT: +5.17 Acc., +3.30 ROC-AUC, and -4.75 EER							
RGB	FFT	85.44	82.61	89.78	86.05	92.64	14.59	**1.08**
	DCT	82.28	81.52	83.47	82.49	89.91	17.72	1.74
	DWT	**92.04**	**92.12**	**91.94**	**92.03**	**97.16**	**7.98**	2.50
	DWT vs. FFT: +6.60 Acc., +4.52 ROC-AUC, and -6.61 EER							

same underlying real-image distribution would further enable the effects of generator shift and source-distribution shift to be studied independently.

Funding. This research received no funding.

Data availability. The code used in this study is publicly available on GitHub.

Competing interests. The authors declare no competing interests.

Appendix A Ablation Study

This appendix reports the ablation experiments used to select the frequency representation, DWT wavelet and boundary mode, and DWT subband configuration. The feature-domain fusion ablation is reported in the main paper in Table 14.

A.1 Selection of Frequency Representation Method

Table A.1 compares FFT, DCT, and DWT representations using RGB and grayscale images. The DRN model is trained and tested on the DFF_A dataset. DWT achieves the highest detection performance for both input representations. For RGB images, DWT achieves 92.04% accuracy and an EER of 7.98%, compared with 85.44% accuracy and 14.59% EER using FFT.

A.2 Selection of DWT Wavelet and Boundary Mode

The Haar, Symlet, and Biorthogonal wavelet families are evaluated with different boundary extension modes. Table A.2 shows that the Symlet wavelet with Reflect boundary extension achieves the highest ROC-AUC of 94.65%, and this configuration is therefore used in SGFF-Net.

A.3 Selection of DWT Subbands

A 2D DWT decomposes an image into four subbands: LL, LH, HL, and HH. To determine the subband configuration used in SGFF-Net, individual subbands (LL and HH), a partial combination (LL+HH), and the complete combination

Table A.2 Ablation of DWT wavelet families and boundary extension modes on DFF_A.

Wavelet	Boundary Mode	ROC-AUC (%)↑
Biorthogonal	Symmetric	91.42
Biorthogonal	Reflect	91.97
Biorthogonal	Antireflect	92.97
Symlet	Symmetric	93.87
Symlet	Reflect	**94.65**
Symlet	Antireflect	93.21
Haar	Symmetric	93.65
Haar	Reflect	94.45
Haar	Antireflect	93.65

(LL+LH+HL+HH) are evaluated under intra-dataset, cross-model, and cross-paradigm settings. Figure A.1 illustrates that combining all four subbands provides the highest accuracy across the evaluated settings. The complete LL+LH+HL+HH configuration is therefore adopted in the final frequency branch.

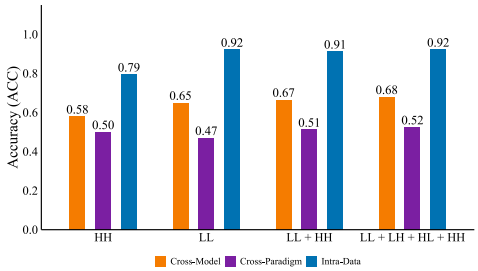


Fig. A.1 Accuracy of different DWT subband configurations under intra-dataset, cross-model, and cross-paradigm evaluation settings.

References

[1] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. & Ortega-Garcia, J. Deepfakes and beyond: a survey of face manipulation and fake detection. *Inf. Fusion* **64**, 131–148 (2020).

[2] Xie, S. *et al.* DeepFake detection in the AIGC era: a survey, benchmarks, and future perspectives. *Inf. Fusion* **127**, 103740 (2026).

[3] Babaei, R., Cheng, S., Duan, R. & Zhao, S. Generative artificial intelligence and the evolving challenge of deepfake detection: a systematic analysis. *J. Sens. Actuator Netw.* **14**, 17 (2025).

[4] World Economic Forum. *The Global Risks Report 2025*. https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf (2025).

[5] Kabir, A. I., Limon, M., Al Fahad, A. & Ridwan, A. Empowering local image generation: harnessing stable diffusion for machine learning and AI. *Inform. Econ.* **28**, 25–38 (2024).

[6] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D. & Lischinski, D. StyleCLIP: text-driven manipulation of StyleGAN imagery. in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 2085–2094 (2021).

[7] Tan, C. *et al.* Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 28130–28139 (2024).

[8] Uhlenbrock, L., Cozzolino, D., Moussa, D., Verdoliva, L. & Riess, C. Did you note my palette? Unveiling synthetic images through color statistics. in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)* 47–52 (2024).

[9] Yu, Y., Ni, R., Li, W. & Zhao, Y. Detection of AI-manipulated fake faces via mining generalized features. *ACM Trans. Multimed. Comput. Commun. Appl.* **18** (2022).

[10] Liu, H. *et al.* Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 772–781 (2021).

[11] Jeong, Y., Kim, D., Ro, Y. & Choi, J. FrePGAN: robust deepfake detection using frequency-level perturbations. in *AAAI Conference on Artificial Intelligence* 1060–1068 (2022).

[12] Jeong, Y. *et al.* BiHPF: bilateral high-pass filters for robust deepfake detection. in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 2878–2887 (2022).

[13] Pontorno, O., Guarnera, L. & Battiato, S. DeepFeatureX-SN: generalization of deepfake detection via contrastive learning. *Multimed. Tools Appl.* (2025). https://doi.org/10.1007/s11042-025-21060-1.

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

[15] Rössler, A. *et al.* FaceForensics++: learning to detect manipulated facial images. in *IEEE/CVF International Conference on Computer Vision (ICCV)* 1–11 (2019).

[16] Nguyen, H. H., Yamagishi, J. & Echizen, I. Capsule-forensics: using capsule networks to detect forged images and videos. in *2019 IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP) 2307–2311 (2019).

[17] He, Z., Chen, P.-Y. & Ho, T.-Y. RIGID: a training-free and generator-agnostic framework for robust AI-generated image detection. *Trans. Mach. Learn. Res.* (2026). https://openreview.net/forum?id=NBkBI2Zjlm.

[18] Liu, Z., Qi, X. & Torr, P. H. S. Global texture enhancement for fake face detection in the wild. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8057–8066 (2020).

[19] Qian, Y., Yin, G., Sheng, L., Chen, Z. & Shao, J. Thinking in frequency: face forgery detection by mining frequency-aware clues. in *European Conference on Computer Vision (ECCV)*, *Lect. Notes Comput. Sci.* **12357**, 86–103 (2020).

[20] Li, J., Xie, H., Li, J., Wang, Z. & Zhang, Y. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 6458–6467 (2021).

[21] Giudice, O., Guarnera, L. & Battiato, S. Fighting deepfakes by detecting GAN DCT anomalies. *J. Imaging* **7**, 128 (2021).

[22] Tan, C., Zhao, Y., Wei, S., Gu, G. & Wei, Y. Learning on gradients: generalized artifacts representation for GAN-generated images detection. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 12105–12114 (2023).

[23] Wu, W., Shi, X., Ai, J. & Wang, X. GRE-Net: a forgery image detection framework based on gradient feature and reconstruction error. *Comput. Vis. Image Underst.* **261**, 104494 (2025).

[24] Singh, S. & Dhumane, A. Unmasking digital deceptions: an integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. *MethodsX* **15**, 103632 (2025).

[25] Sharafudeen, M. & Chandra, S. S. V. Frequency forensics for deep fake face detection using dual residual networks. *Multimed. Tools Appl.* **84**, 44063–44088 (2025).

[26] Bammey, Q. Synthbuster: towards detection of diffusion model generated images. *IEEE Open J. Signal Process.* **5**, 1–9 (2024).

[27] Guan, L., Liu, F., Zhang, R., Liu, J. & Tang, Y. MCW: a generalizable deepfake detection method for few-shot learning. *Sensors* **23**, 8763 (2023).

[28] Hasanaath, A. A., Luqman, H., Katib, R. & Anwar, S. FSBI: deepfake detection with frequency enhanced self-blended images. *Image Vis. Comput.* **154**, 105418

(2025).

[29] Uddin, M., Fu, Z. & Zhang, X. Deepfake face detection via multi-level discrete wavelet transform and vision transformer. *Vis. Comput.* **41**, 7049–7061 (2025).

[30] Baru, L. B., Boddeda, R., Patel, S. A. & Gajapaka, S. M. Wavelet-driven generalizable framework for deepfake face forgery detection. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* 769–778 (2024).

[31] Yu, N., Davis, L. S. & Fritz, M. Attributing fake images to GANs: learning and analyzing GAN fingerprints. in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 7556–7566 (2019).

[32] Marra, F., Gragnaniello, D. & Poggi, G. Do GANs leave artificial fingerprints? in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* 506–511 (2019).

[33] Dasgupta, S., Badal, K., Chittam, S., Alam, M. T. & Roy, K. Attention-enhanced CNN for high-performance deepfake detection: a multi-dataset study. *IEEE Access* **13**, 101980–101993 (2025).

[34] Ojha, U., Li, Y. & Lee, Y. J. Towards universal fake image detectors that generalize across generative models. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 24480–24489 (2023).

[35] Corvi, R. *et al.* On the detection of synthetic images generated by diffusion models. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1–5 (2023).

[36] Sha, Z., Li, Z., Yu, N. & Zhang, Y. DE-FAKE: detection and attribution of fake images generated by text-to-image generation models. in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS)* 3418–3432 (2023).

[37] Marra, F., Gragnaniello, D., Cozzolino, D. & Verdoliva, L. Detection of GAN-generated fake images over social networks. in *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* 384–389 (2018).

[38] Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. in *2017 IEEE International Conference on Computer Vision (ICCV)* 2242–2251 (2017).

[39] Ameer, N. H. A., Al-Taee, M., Hussain, A.-S. T., Hussian, A. & Ali, H. The impact of data augmentation on deepfake detection: a comparative study of CNN and transformer models. in *2025 IEEE 4th International Conference on Computing*

and Machine Intelligence (ICMI) 1–6 (2025).

[40] Dang, H., Liu, F., Stehouwer, J., Liu, X. & Jain, A. K. On the detection of digital face manipulation. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 5781–5790 (2020). Available at http://cvlab.cse.msu.edu/dffd-dataset.html.

[41] Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. in *International Conference on Learning Representations (ICLR)* (2021).

[42] Nichol, A. & Dhariwal, P. Improved denoising diffusion probabilistic models. in *Proceedings of the International Conference on Machine Learning (ICML)*, *Proc. Mach. Learn. Res.* **139**, 8162–8171 (2021).

[43] Reddy, M. D. M., Basha, M. S. M., Hari, M. M. C. & Penchalaiah, M. N. Dall-e: creating images from text. *Dogo Rangsang Res. J.* **8**, 71–75 (2021).

[44] Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G. & Verdoliva, L. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. in *IEEE International Conference on Multimedia and Expo (ICME)* 1–6 (2021).

[45] Chen, T. *et al.* Masked conditional diffusion model for enhancing deepfake detection. in *2024 International Joint Conference on Neural Networks (IJCNN)* 1–7 (2024).

[46] Khan, S. A. & Dang-Nguyen, D.-T. Deepfake detection: analyzing model generalization across architectures, datasets, and pre-training paradigms. *IEEE Access* **12**, 1880–1908 (2024).

[47] Song, H., Huang, S., Dong, Y. & Tu, W.-W. Robustness and generalizability of deepfake detection: a study with diffusion models (2023). Preprint at https://arxiv.org/abs/2309.02218.

[48] Chen, Z. *et al.* DiffusionFace: towards a comprehensive dataset for diffusion-based face forgery analysis (2024). Preprint at https://arxiv.org/abs/2403.18471v1.

[49] Rothe, R., Timofte, R. & Van Gool, L. DEX: deep EXpectation of apparent age from a single image. in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)* 252–257 (2015).

[50] Xia, W., Yang, Y., Xue, J.-H. & Wu, B. TediGAN: text-guided diverse face image generation and manipulation. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2256–2265 (2021).

[51] Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 4401–4410 (2019).

[52] Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. in *Advances in Neural Information Processing Systems (NeurIPS)* **33**, 6840–6851 (2020).

[53] Liu, L., Ren, Y. & Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. in *International Conference on Learning Representations (ICLR)* (2022).

[54] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10684–10695 (2022).

[55] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016).

[56] Li, Y., Yang, X., Sun, P., Qi, H. & Lyu, S. Celeb-DF: a large-scale challenging dataset for DeepFake forensics. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 3207–3216 (2020).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SGFFNetMethodology.pdf
- DWTBlock.pdf
- DRN.pdf
- misclassifiedexamples.pdf