2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

FSSNet: Frequency-Spatial Synergy Network for Universal Deepfake Detection

Zepeng Su[1] and Yin Chen[1]*

[1] School of Artificial Intelligence, South China Normal University, Guangdong, China
suzep@foxmail.com
*Corresponding author

Abstract. In this paper, our aim is to develop a detector capable of effectively identifying previously unseen deepfake images, even with limited training data. Existing deepfake detection methods predominantly focus on single modality. For instance, frequency-domain approaches leverage Fourier transforms to capture frequency information, while spatial-domain methods utilize convolutional networks to extract visual features. However, relying on a single modality limits the ability to capture diverse feature types, resulting in poor generalization. To overcome this limitation, we propose a dual-stream network, FSSNet, which integrates the Scale-aware Bidirectional Cross Attention (SBCA) module and the Adaptive Feature Fusion (AFF) module for comprehensive and dynamic multi-modal feature fusion. Experimental results on deepfake images generated by eight unseen GAN models and ten unseen diffusion models demonstrate the superior performance of FSSNet, showcasing its robust generalization capability.

Keywords: Deep learning, Deepfake detection, Multi-modal fusion.

1 Introduction

In recent years, the rapid advancement of deep learning techniques has facilitated the emergence of deepfake technology, While these innovations present exciting opportunities for creative expression, they also pose significant challenges, particularly in the realm of misinformation and digital security. As deepfake content becomes increasingly sophisticated, An effective and generalizable detector is paramount.

Wang et al.[1] first revealed universal artifacts in CNN-generated images, demonstrating that classifiers trained on a single model can generalize to unseen models. Experiments on various GANs (e.g., ProGAN, StyleGAN, BigGAN) showed consistent artifacts across datasets and generation methods. This work established a solid foundation for deepfake detection, providing a strong benchmark and theoretical support. Building on this, subsequent studies focused on enhancing generalization, primarily through single-modal feature extraction in either the frequency or spatial domain.

However, single-modal methods have significant limitations. Frequency-domain approaches, while extracting spectral features via Fourier transforms, yet neglect spatial context like structures and textures, reducing their effectiveness on complex visuals. Conversely, spatial-domain methods focus on visual features but fail to capture frequency distribution biases. These shortcomings limit their ability to generalize across models and datasets, making them inadequate for handling the complexities of diverse deepfake generation techniques.

To overcome the limitations of single-modal methods, we propose FSSNet (Frequency-Spatial Synergy Network). As shown in Fig.1, unlike single-modal methods, FSSNet simultaneously focuses on both spatial and frequency domains, using two independent backbones to extract features from spatial and frequency information. Then, carefully designed information interaction and fusion modules are used to achieve complementary enhancement of the features, fully leveraging their complementary information for classification.

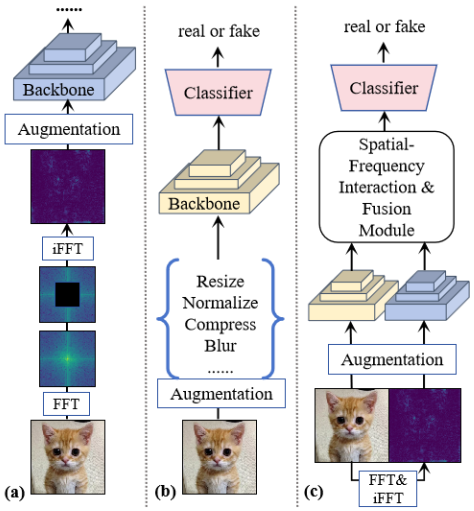

Fig. 1. Our Multi-Modal Method vs. Existing Single-Modal Methods}. (a) and (b) represent the general frameworks of frequency-based and spatial-based detectors, respectively, while our method (c) effectively aggregates and fully utilizes information from both domains.

FSSNet includes two key modules: Scale-Aware Bidirectional Cross-Attention (SBCA), which captures multi-granular information for deep cross-modal interaction, and Adaptive Feature Fusion (AFF), which dynamically combines spatial and frequency features using channel and spatial attention with a learnable factor. This multi-modal approach overcomes the limitations of single-modal methods, enabling robust performance across GAN-generated and unseen generative models, including diffusion models, with strong generalization capabilities.

The main contributions of our work are as follows:

● We propose the Frequency-Spatial Synergy Network (FSSNet), which enables complementary integration of spatial and frequency domain information. The SBCA module facilitates cross-modal interaction and enhancement, while the AFF module dynamically fuses the two types of modal features.
● We are the first to introduce DINOv2[2] as the backbone for deepfake detection, leveraging its pre-training on large-scale data to enhance feature extraction and representation quality. We fine-tune DINOv2 to better adapt it to the specific requirements of our deepfake detection task.
● To evaluate the effectiveness and generalization capability of our method, we performed experiments on eight previously unseen GAN models and ten previously unseen Diffusion Models (DMs). Our method achieved an accuracy (Acc.) of 94.8 and an average precision (A.P.) of 97.0 on the GAN test set, and an Acc. of 92.6 and an A.P. of 97.2 on the DM test set, both outperforming the current state-of-the-art.

2 Related Work

2.1 Generative Models

With rapid advancements in generative models for computer vision and image synthesis, their applications in deepfake generation have become increasingly widespread.

Generative Adversarial Networks (GANs)[3], introduced by Goodfellow et al. in 2014, mark a significant breakthrough in generative model research. GANs comprise a generator and a discriminator trained adversarially, enabling the generator to produce images closely resembling real data distributions. Since their inception, a variety of GAN variants have emerged, each introducing novel techniques to enhance the stability, diversity, and quality of the generated images.

In recent years, Diffusion Models (DMs) [4] have emerged as a new paradigm in generative modeling. By progressively adding noise to data and learning the reverse process, DMs generate high-quality samples. Notable examples of DMs include DALL-E[17], a model that generates highly detailed images from textual descriptions, while guided diffusion models, such as GLIDE[18], leverage additional conditioning information (e.g., text or images) to guide the generative process, leading to more controlled and targeted image outputs. Compared to GANs, DMs offer superior generation stability and diversity.

2.2 Spatial-Based Detection

Spatial-based methods primarily focus on detecting visual artifacts and inconsistencies at the pixel level. For example, Wang et al.[1] proposed a detection method that trains a ResNet-based classifier to leverage common artifacts found in CNN-generated images. Ju et al.[20] developed a two-branch model that improves the detection of AI-generated images by combining global and local features. Tan et al.[5] introduced a method that transforms images into gradients to extract generalized artifact features from a pretrained CNN model. Similarly, Ojha et al.[6] proposed a universal fake image detection method utilizing the feature space of the large pretrained vision-language

model CLIP[19]. This method uses a nearest-neighbor classification approach, comparing the target image with known real images to assess its authenticity. By leveraging pretrained model features, this approach significantly simplifies the training process.

2.3 Frequency-Based Detection

Frequency-based methods analyze image features in the frequency domain to detect anomalies that are often imperceptible in the spatial domain. For example, Frank et al.[7] conducted the first systematic analysis of the spectral characteristics of GAN-generated images, revealing significant artifacts in the frequency domain for all such images. These frequency artifacts can be effectively leveraged to construct efficient deepfake detectors. Durall et al.[8] highlighted the failure of many generative models to replicate real image spectral properties, enabling detection. Methods such as Bi-HPF[9] amplify high-frequency artifacts to enhance cross-domain detection performance, while FrePGAN[10] mitigates frequency artifacts via perturbation maps, enhancing generalization across models. Tan et al.'s FreqNet[11] further employs convolution on phase and amplitude spectra, effectively capturing high-frequency artifacts and boosting cross-domain performance.

3 Method

3.1 Problem Setup

Our study aims to develop a deepfake detection model with robust generalization capabilities. Specifically, we train the model D using images generated by a single generator, ProGAN, and evaluate its performance in detecting deepfake images produced by various unseen generators. The detection model Dlearns to classify images as real or fake by extracting features from both real images and ProGAN-generated images. Formally, let $x \in R^{H\times W\times C}$ denote an input image, where H, W and C are its height, width, and number of channels. The objective is to learn a mapping: $\mathcal{D}: x \in R^{H\times W\times C} \rightarrow \{0,1\}$, where:

- $\mathcal{D}(x = 0)$ indicates a real image,
- $\mathcal{D}(x = 1)$ indicates a ai-generated image.

To optimize the model, we minimize a loss function $\mathcal{L}(\mathcal{D})$ with respect to the model parameters θ,where y represents the ground truth label:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\mathcal{D}(x;\theta), y). \tag{1}$$

Testing is conducted on images generated by multiple unseen models. Define the test set as $T_{test} = \{G_1, G_2, \dots, G_n\}$, where each G_i represents a distinct generative model. Our goal is for the trained model $\mathcal{D}$ to satisfy the following conditions:

$$\mathcal{D}(x_{G_i}) \approx y_{G_i}, (G_i \in \{G_1, G_2, \dots, G_n\}). \tag{2}$$

3.2 Overall Architecture

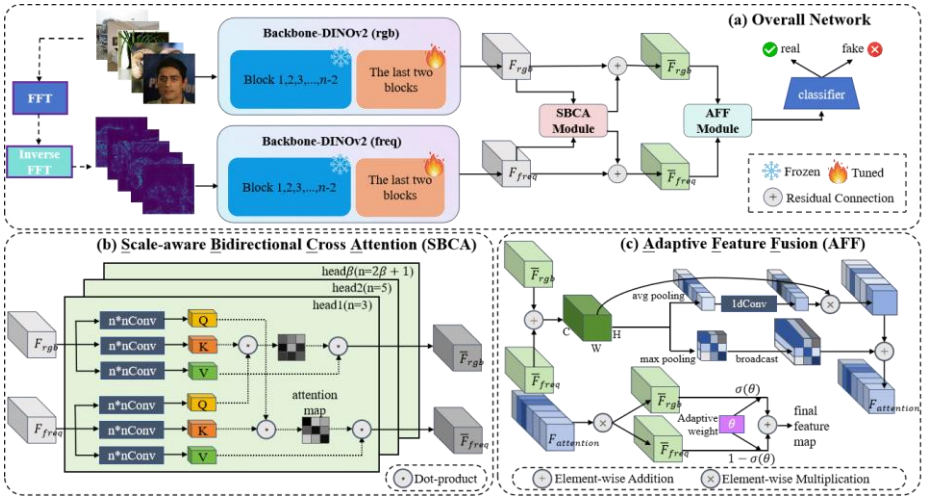

Fig. 2. Overall Architecture for Generalizable Deepfake Detection. (a) Overall Network. DINOv2 serves as the backbone to separately extract spatial and frequency domain features. The fused feature maps are obtained through the SBCA and AFF modules, followed by a linear layer classifier for deepfake detection. (b) Scale-aware Bidirectional Cross Attention. Leveraging multi-head attention, each head captures interaction information at different scales, enabling bidirectional interaction between rgb and frequency features. (c) Adaptive Feature Fusion. This module concurrently incorporates spatial and channel attention, employing learnable weights for dynamic feature fusion.

To achieve robust and highly generalized deepfake detection, we propose the Frequency-Spatial Synergy Network (FSSNet), which dynamically and comprehensively fuses spatial and frequency domain information. The overall framework is illustrated in Fig.2.

Given an input image $I \in R^{H\times W\times C}$, two copies are generated. One copy, I_s is directly fed into a backbone designed to extract spatial domain features. The other copy, I_f , undergoes a series of transformations in the frequency domain. First, the Fast Fourier Transform(FFT)[12] is applied to obtain the frequency domain representation F:

$$F = FFT(I_f), \tag{3}$$

where $F \in \mathbb{C}^{H\times W}$ represents the frequency domain features of the image. Next, frequency band selection is performed by applying a mask M to retain specific frequency components:

$$F' = F \odot M. \tag{4}$$

Here, $\odot$ denotes element-wise multiplication, and $M \in \{0,1\}^{H\times W}$ is a binary mask used to select low-frequency, mid-frequency, or high-frequency components. After selecting the frequency bands, the filtered frequency representation F' is transformed back into the spatial domain via the Inverse Fast Fourier Transform($iFFT$):

$$I'_f = iFFT(F'), \tag{5}$$

where, I'_f is subsequently fed into a separate backbone to extract frequency domain features. These two backbones are trained independently without sharing weights to better capture the specific characteristics of each domain.

Finally, the spatial and frequency features are passed through two carefully designed modules: the Scale-aware Bidirectional Cross-Attention (SBCA) module and the Adaptive Feature Fusion (AFF) module. These modules facilitate efficient cross-domain feature interaction and dynamic fusion. The fused features are then classified by a linear classifier to generate the detection result.

3.3 Scale-aware Bidirectional Cross Attention

To facilitate effective interaction between spatial and frequency domain information, we propose the Scale-aware Bidirectional Cross-Attention (SBCA) module. This module enables bidirectional information flow between the two domains, ensuring a more comprehensive capture of their associations and complementary characteristics.

- Spatial-to-Frequency Attention: Spatial domain features serve as Query (Q), while frequency domain features act as Key(K) and Value (V). Cross-attention is performed as follows:

$$Attn_{s2f}(Q_s, K_f, V_f) = softmax\left(\frac{Q_s {K^T}_f}{\sqrt{d}}\right)V_f. \tag{6}$$

- Frequency-to-Spatial Attention: Conversely, frequency domain features are used as Query(Q), and spatial domain features as Key(K) and Value (V):

$$Attn_{f2s}(Q_f, K_s, V_s) = softmax\left(\frac{Q_f {K^T}_s}{\sqrt{d}}\right)V_s. \tag{7}$$

This bidirectional design enables the model to dynamically integrate spatial and frequency features in both directions. In addition, to capture information at various granularities, we adopt a multi-head attention mechanism with multi-scale convolutional projections for generating (Q), (K), and (V). In each attention head, convolutions with different kernel sizes are applied to achieve scale-aware feature extraction. The convolution kernel size for the β-th head is defined as:$k_\beta = 2\,\beta + 1$, The multi-scale convolutional projection is formulated as:

$$Q^\beta = Conv(X_Q; k_\beta), \tag{8}$$

$$K^\beta = Conv(X_Q; k_\beta), \tag{9}$$

$$V^\beta = Conv(X_Q; k_\beta), \tag{10}$$

where $Conv(X; k_\beta)$ denotes a convolution operation on the input feature map X with a kernel size of k_β, and $\beta \in \{1,2,3,\ldots,H\}$ represents the index of the attention head. H is the total number of heads.

The feature maps F_{rgb} and F_{freq} obtained from the backbone are processed through the SBCA module, resulting in the enhanced and interacted feature maps $\bar{F}_{rgb}$ and $\bar{F}_{freq}$. By using different kernel sizes across attention heads, the model can achieve fine-

grained and coarse-grained interactions between spatial and frequency domain features, thereby enhancing feature representations.

3.4 Adaptive Feature Fusion

In the Adaptive Feature Fusion(AFF) module, we fully exploit both the spatial and channel information of the features by jointly utilizing spatial attention and channel attention to generate the attention map $F_{attention}$. This attention map is then applied to both the spatial domain features $\bar{F}_{rgb}$ and the frequency domain features $\bar{F}_{freq}$ through element-wise multiplication, resulting in the final feature map F_{final} ,which is used as the input for the classifier.

Specifically, a learnable parameter θ is introduced, dynamically guiding the fusion of the two domains. θ is a adaptive weight during the training process, constrained within the range [0,1] by the sigmoid function(σ), and initialized to 0.5 after applying the sigmoid function. This allows the model to dynamically adjust the relative importance of spatial and frequency domain features during training, enabling an adaptive and balanced fusion mechanism. The final feature map F_{final} is computed as:

$$F_{final} = \sigma(\theta) \cdot \left(F_{attention} \odot \bar{F}_{rgb}\right) + \left(1 - \sigma(\theta)\right) \cdot \left(F_{attention} \odot \bar{F}_{freq}\right) \quad (11)$$

4 Experiments

4.1 Training Dataset

We adopt a setup consistent with prior baselines (e.g., Wang et al.[1], Ojha et al.[6], and Tan et al.[11]), training exclusively on a single generative model, specifically using ProGAN-generated images to train the detector, but testing its detection capabilities on various unseen GAN and diffusion models. The training dataset is derived from the ForenSynths dataset[1] (Wang et al.), with each class containing 18,000 real images and their corresponding ProGAN-generated images. In alignment with the benchmarking methodology, we adopt a 4-class training configuration, wherein the detection model is trained on images spanning four distinct categories: horse, chair, car, and cat.

4.2 Testing Dataset

To evaluate the generalization ability of our detection model in real-world scenarios, our test set includes a variety of GAN and Diffusion generative models. Specifically, the GAN test set was sourced from the ForenSynths dataset, while the Diffusion model test set was sourced from the Ojha's and DIRE datasets.

- The 8 GANs test models from ForenSynths[1]: ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, and Deepfake.
- The 10 Diffusion test models from Ojha's[6] and DIRE[13]: PNDM, DALL-E, VQ-Diffusion, Guided, LDM, and Glide. For the LDM and Glide models, three different generation parameters were used for each.

4.3 Implementation Details

We employed the pre-trained DINOv2 (ViT-B/14) as the backbone, fine-tuning only the last two blocks. Input images were cropped to 252x252 to align with the backbone's requirements. The training process utilized a batch size of 64 and an initial learning rate of 4e-06. We adopted the AdamW[15] optimizer and applied cosine annealing for dynamic learning rate adjustment. The model was trained for 50 epochs, with an early stopping mechanism: training was halted if performance failed to improve over 6 consecutive epochs, thereby conserving time and computational resources. Our method was implemented in PyTorch[16] and ran on an NVIDIA GeForce RTX 3090 GPU. Consistent with established baselines, we evaluated the detection model using accuracy (Acc.) and average precision (A.P.) as performance metrics.

4.4 Evaluating Generalization on GANs Dataset

Table 1. Generalization performance on the unseen GANs (Generative Adversarial Networks) dataset.

Methods	Input	Pro GAN	Style-GAN	Style-GAN2	Big-GAN	Cy-cleGAN	Star-GAN	Gau-GAN	Deep-fake	Mean
		Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.
		A.P.	A.P.	A.P.	A.P.	A.P.	A.P.	A.P.	A.P.	A.P.
Wang[1]	Rgb	50.4	50.4	68.2	50.2	50.0	50.0	50.3	50.1	52.5
(CVPR 2020)		63.8	79.4	94.7	61.3	52.9	48.2	67.6	51.5	64.9
Durall[8]	Freq	85.1	59.2	70.4	57.0	66.7	99.8	58.7	53.0	68.7
(CVPR 2020)		79.5	55.2	63.8	53.9	61.4	99.6	54.8	51.9	65.0
BiHPF[9]	Freq	90.7	76.9	76.2	84.9	81.9	94.4	69.5	54.4	78.6
(WACV 2022)		86.2	75.1	74.7	81.7	78.9	94.4	78.1	54.6	77.9
SelfBlend[14]	Rgb	58.8	50.1	48.6	51.1	59.2	74.5	59.2	93.8	61.9
(CVPR 2022)		65.2	47.7	47.4	51.9	65.3	89.2	65.5	99.3	66.4
FreP-GAN[10]	Freq	99.0	80.7	84.1	69.2	71.1	99.9	60.3	70.9	79.4
(AAAI 2022)		99.9	89.6	98.6	71.1	74.4	100.0	71.7	91.9	87.2
Lgrad[5]	Rgb	99.0	94.8	96.0	82.9	85.3	99.6	72.4	58.0	86.1
(CVPR 2023)		100.0	99.9	99.9	90.7	94.0	100.0	79.3	67.9	91.5
Ojha[6]	Rgb	99.7	89.0	83.9	90.5	87.9	91.4	89.9	80./	89.1
(CVPR 2023)		100.0	98.7	98.4	99.1	99.8	100.0	100.0	90.2	98.3
FreqNet[11]	Freq	99.6	90.2	88.0	90.5	95.8	85.7	93.4	88.9	91.5
(AAAI 2024)		100.0	99.7	99.5	96.0	99.6	99.8	98.6	94.4	**98.5**
FSSNet(our)	Rgb&Freq	99.9	93.0	95.3	99.6	97.7	98.3	99.8	74.6	**94.8**
		100.0	98.2	98.8	99.9	99.7	99.8	100.0	79.7	97.0

As shown in Table 1, we evaluate the generalization capability of our model on 8 datasets generated by various GANs, where the **Bold** indicates the best results.

We compare our proposed method with eight baseline and state-of-the-art (SOTA) methods:Wang's method, Durall's method, BiHPF, FrePGAN, LGrad, SelfBlend, Ojha's method, and FreqNet. The experimental results of the baseline and SOTA methods are taken from their respective papers or prior reproductions.

The results demonstrate that our proposed method, FSSNet, surpasses SOTA methods, achieving superior accuracy (Acc.) and average precision (A.P.), achieves an Acc. score of 94.8 and an A.P. score of 97.0. Notably, our method, which integrates spatial and frequency domain information, demonstrates stronger generalization capability compared to SOTA single-modal approaches such as Ojha's method and FreqNet, surpasses Ojha's and FreqNet by 5.7% and 3.3% in Acc., respectively. This highlights that by effectively fusing spatial and frequency domain features, a detector trained on a single dataset can be successfully generalized to detect ai-generated images by unseen GANs.

4.5 Evaluating Generalization on DMs Datasets

Table 2. Generalization performance on the unseen DMs (Diffusion Models) dataset.

Methods	PNDM	DALL-E	VQ-Diffusion	LDM			Guided	Glide			Mean
				100	200	200-cfg		100-10	100-27	50-27	
	Acc. A.P.	Acc. A.P.	Acc.A.P.	Acc. A.P.	Acc. A.P.	Acc. A.P.	Acc. A.P.	Acc. A.P.	Acc. A.P.	Acc. A.P.	Acc. A.P.
Wang[1] (CVPR 2020)	50.8 90.3	51.8 61.3	50.0 71.0	51.9 63.7	52.0 64.5	51.6 63.1	54.9 66.6	53.3 72.9	53.0 71.3	52.4 70.1	50.8 90.3
Durall[8] (CVPR 2020)	44.5 47.3	55.9 58.0	38.6 38.3	62.0 62.9	61.7 61.7	58.4 58.5	40.6 42.3	54.9 52.3	48.9 46.9	51.7 51.8	44.5 47.3
SelfBlend [14] (CVPR 2022)	48.2 48.2	52.4 51.6	77.2 82.7	53.0 54.0	52.6 51.9	51.9 52.6	58.3 63.4	58.8 63.2	59.4 64.1	64.2 68.3	57.6 60.0
Lgrad[5] (CVPR 2023)	69.8 98.5	88.5 97.3	96.3 100.0	94.8 99.2	94.2 99.1	95.9 99.2	86.6 100.0	89.4 94.9	87.4 93.2	90.7 95.1	89.4 97.7
Ojha[6] (CVPR 2023)	75.3 92.5	89.5 96.8	83.5 97.7	90.5 97.0	90.2 97.1	77.3 88.6	75.7 85.1	90.1 97.0	90.7 97.2	91.1 97.4	85.4 94.6
FreqNet[11](AAAI 2024)	85.2 99.9	97.2 99.7	100.0 100.0	97.8 99.9	97.4 99.9	97.2 99.9	67.2 75.4	87.8 96.0	84.4 95.6	86.6 95.8	90.1 96.2
FSSNet(our)	97.9 99.8	96.2 99.1	97.0 99.5	97.2 99.6	97.6 99.6	89.8 96.0	88.9 95.0	86.9 94.3	88.0 95.2	86.7 93.9	**92.6** **97.2**

As shown in Table 2, we evaluate the generalization capability of our model in more challenging evaluation scenarios by comparing our method with existing detection approaches on 10 previously unseen diffusion model datasets.(Input are the same as in Table1). The baseline and SOTA methods compared include Wang's method, Durall's method, SelfBlend, LGrad, Ojha's method, and FreqNet. Notably, as FrePGAN and BiHPF is not open-sourced, we could not obtain its results on diffusion models. Furthermore, since FreqNet has not been evaluated for generalization on diffusion models in their paper, we reproduce its results using the official pre-trained model and code[11]. The results for the other methods are drawn from their respective papers or prior reproductions.

Surprisingly, similar to the previous results, our FSSNet method demonstrates stronger generalization capability on previously unseen diffusion model datasets compared to single-modal SOTA methods such as Ojha's method and FreqNet. Specifically, FSSNet outperforms Ojha's method and FreqNet by 7.2% and 2.5% in Acc., respectively. These results further validate the effectiveness of our approach, showing its ability to generalize well to unseen diffusion models.

5 Ablation study

To evaluate the impact of different frequency bands on model performance, we conducted ablation experiments by retaining low, mid, and high-frequency components separately and assessing detection accuracy (Acc.) and average precision (A.P.). As shown in Table 3, the low-frequency components achieved the best results for GANs (94.78% Acc., 97.02% A.P.). For diffusion models (DMs), the differences between low and mid-frequency components were negligible, leading us to conclude that low-frequency information is more indicative of generative model artifacts and is suitable for providing supplementary information for detection.

Table 3. The table presents the average accuracy(Acc.) and average precision(A.P.) of GANs and DMs dataset in different frequency bands.

Generative Model	Frequency Bands		
	Low	Mid	High
	Acc./A.P.	Acc./A.P.	Acc./A.P.
GANs	94.78/97.02	94.24/96.51	94.42/96.66
DMs	92.54/97.18	92.58/97.18	92.30/97.05
Average	**93.66/97.10**	93.41/96.85	93.36/96.86

To further evaluate the contributions of multimodal information, the Scale-aware Bidirectional Cross Attention (SBCA) module, and the Adaptive Feature Fusion (AFF) module, we conducted ablation experiments. As shown in Table 4, using only single-modal information leads to a significant decline in model performance, highlighting the importance of integrating multimodal information. Additionally, when multimodal information is combined, the removal of either the SBCA or AFF module results in a certain degree of performance degradation. This further underscores the critical roles

of these modules in dynamically fusing spatial and channel information, demonstrating their complementary contributions to enhancing the generalization capability of the detection model.

Table 4. The table presents the average accuracy(Acc.) and average precision(A.P.) of our proposed method under the following conditions: single information domain(only rgb & only freq way), without the SBCA module, and without the AFF module.

Generative Model	Rgb info	Freq info	SBCA	AFF	Mean Acc./A.P.	Degradation Acc./A.P.
GANs	✓				89.78/91.52	↓ 5.00/5.50
		✓			89.14/91.05	↓ 5.64/5.97
	✓	✓		✓	93.50/95.77	↓ 1.28/1.25
	✓	✓	✓		92.88/95.20	↓ 1.90/1.82
	✓	✓	✓	✓	94.78/97.02	-
DMs	✓				88.35/93.51	↓ 4.23/3.67
		✓			87.94/92.88	↓ 4.64/4.30
	✓	✓		✓	90.98/96.39	↓ 1.60/0.79
	✓	✓	✓		91.92/97.30	↓ 0.66/-0.12
	✓	✓	✓	✓	92.58/97.18	-

6 Conclusion

We presents a novel dual-stream network, FSSNet, designed to enhance the generalization capability of deepfake detection. FSSNet integrates several carefully designed modules that effectively capture cross-modal interactions and dynamically fuse multi-modal features, significantly improving detection performance. Experimental results demonstrate the potential of multi-modal feature fusion in addressing sophisticated deepfake techniques and provide guidance for improving model generalization in more complex generative scenarios.

References

1. Wang, Sheng-Yu, et al. "CNN-generated images are surprisingly easy to spot... for now." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
2. Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." *arxiv preprint arxiv:2304.07193* (2023).
3. Goodfellow, Ian J., et al. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).
4. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

5. Tan, Chuangchuang, et al. "Learning on gradients: Generalized artifacts representation for gan-generated images detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
6. Ojha, Utkarsh, Yuheng Li, and Yong Jae Lee. "Towards universal fake image detectors that generalize across generative models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
7. Frank, Joel, et al. "Leveraging frequency analysis for deep fake image recognition." *International conference on machine learning*. PMLR, 2020.
8. Durall, Ricard, Margret Keuper, and Janis Keuper. "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
9. Jeong, Yonghyun, et al. "Bihpf: Bilateral high-pass filters for robust deepfake detection." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.
10. Jeong, Yonghyun, et al. "Frepgan: robust deepfake detection using frequency-level perturbations." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. No. 1. 2022.
11. Tan, Chuangchuang, et al. "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 5. 2024.
12. Cooley, James W., and John W. Tukey. "An algorithm for the machine calculation of complex Fourier series." *Mathematics of computation* 19.90 (1965): 297-301.
13. Zhang, Yichi, and Xiaogang Xu. "Diffusion noise feature: Accurate and fast generated image detection." *arxiv preprint arxiv:2312.02625* (2023).
14. Shiohara, Kaede, and Toshihiko Yamasaki. "Detecting deepfakes with self-blended images." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
15. Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." *arxiv preprint arxiv:1711.05101* (2017).
16. Paszke, A. "Pytorch: An imperative style, high-performance deep learning library." *arxiv preprint arxiv:1912.01703* (2019).
17. Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *International conference on machine learning*. Pmlr, 2021.
18. Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." *arxiv preprint arxiv:2112.10741* (2021).
19. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021.
20. Ju, Yan, et al. "Fusing global and local features for generalized ai-synthesized image detection." 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022.