J. Appl. Numer. Optim. 7 (2025), No. 1, pp. 97-111
Available online at http://jano.biemdas.com
https://doi.org/10.23952/jano.7.2025.1.07

EDFM: AN ENHANCED DUAL-BRANCH FUSION MODEL FOR FACE DEEPFAKE DETECTION

ZHENGXIN ZHOU, JUNMEI SUN, XIUMEI LI*

Department of Information Science and Technology, Hangzhou Normal University, Hangzhou, China

Abstract. Face deepfake technology brings serious security risks such as privacy leakage, false information dissemination, and network fraud, which need to be widely concerned and prevented. In recent years, many detection methods were proposed, among which enhancing the robustness and generalization ability of the model has always been an important topic. In this paper, we propose a novel enhanced dual-branch fusion model to improve the robustness and generalization ability of CNN-based face deepfake detector. Our method begins by enhancing the RGB high-frequency noise in the face image to extract its abnormal features, and then performs preservation fusion. Specifically, we use a deep separable convolution module to improve the model performance when extracting image features. When extracting noise features, we use a selective kernel module to adaptively extract more representative noise features by dynamically adjusting the convolution kernel. In addition, we specially design a multi-scale channel spatial attention fusion module to effectively fuse the feature information of each part, thereby reducing model overfitting and enhancing the robustness and generalization ability of the model. Finally, through comprehensive evaluation on several benchmark datasets, it is confirmed that our method has significantly improved robustness and generalization.

Keywords. Deep face forgery detection; Dual-branch network; Multi-scale fusion.
2020 Mathematics Subject Classification. 68U07, 68W10.

1. INTRODUCTION

Face deepfake technology [1] uses deep learning models to generate highly realistic fake face images and videos. With the advancement of forgery techniques, these synthetic contents are almost visually indistinguishable from real images, posing serious threats to personal privacy, public opinion manipulation, and public safety [2, 15, 29]. To mitigate such risks, deepfake detection technology has become an essential research direction [23]. Current detection methods can be divided into two main categories: feature extraction-based methods and deep learningba-sed methods [10, 24]. The former identifies forgery by extracting manually designed image features, such as abnormal facial textures and inconsistent lighting [4]. In contrast, deep learningb-ased methods [16] use models like convolutional neural networks (CNNs) to automatically learn feature representations that distinguish between real and fake content, achieving high detection accuracy [34]. Despite of the proposed methods, deepfake detection still faces

*Corresponding author.
E-mail address: lixiumei@hznu.edu.cn (X. Li).
Received 11 January 2025; Accepted 9 March 2025; Published online 20 March 2025.

several challenges, especially in detecting forged datasets generated by different forgery methods, where robustness is clearly lacking. Additionally, some studies [3, 19] found that, due to the model's insufficient accuracy in extracting and selecting features, generalization across datasets is limited. Therefore, further enhancing the robustness and generalization capability of models in cross-dataset detection algorithms is crucial for information security and robustness [35].

Recently, to address the issue of insufficient generalization capability, researchers begun exploring more refined detection methods, including dual-branch fusion strategies [37] , transfer learning [30], domain adaptation and multi-task learning [36], to enhance model generalization. The dual-branch fusion method combines information from different feature extraction branches, allowing for a more comprehensive capture of multi-level features in forged images, thereby improving detection accuracy and robustness [40]. This strategy typically integrates information from both spatial and frequency domains, with two independent branches handling detailed and global features, respectively, and a fusion module synthesizing them to assess the likelihood of forgery [25]. Additionally, some studies attempted to obtain useful information for face deepfake detection from the frequency domain, such as using Discrete Cosine Transform(DCT), steganalysis features, and Fourier transform [13]. However, these methods have not fully leveraged the interaction and extraction of features from normal images, failing to thoroughly exploit the image data. Previous dual-branch recognition structures [12, 31] applied the same processing to both the RGB branch and the high-frequency noise branch by using Spatial Rich Model (SRM) [6]. Upon analyzing the high-frequency noise suppression areas and RGB feature maps, it was found that training is overly focused on regions irrelevant to forgery detection, leaving room for improvement in generalization. In this paper, our goal is to further enhance the generalization and optimize the robustness of dual-branch face forgery detectors. We propose an optimized solution for cross-dataset face forgery detection models. To more adaptively utilize image noise and conventional image features, and to more reasonably fuse the two features for improved robustness and generalization, we designed a combination of three modules, each acting on the dual branches and the fusion process. The first is a Depthwise Separable Convolutional Module (DSCM), which reduces the performance overhead in the RGB branch while ensuring stability. The second is a Selective Kernel Module (SKM) [14], which uses a dyna-mic selection mechanism to adjust the receptive field size adaptively in the noise extraction branch. After constraining the features from the two branches, we apply our proposed multi-scale Branch Fusion Module (BFM), which performs multi-scale fusion of the enhanced dual branches across spatial and channel dimensions, further weighting the core features from both branches.

Our contributions are summarized as follows:

• We propose a novel enhanced dual-branch fusion mo-del to improve the robustness and generalization capability of CNN-based face deepfake detectors.

• Based on the functional analysis of the RGB and hig-h-frequency noise branches, we introduce the DSCM and SKM modules to optimize the corresponding branches. In the RGB branch, we reduce performance overhead while maintaining feature extraction capability, and in the high-frequency noise branch, we employ DSCM to more precisely expose tampered regions.

• During the dual-branch fusion process, we propose a multi-scale DFM that extracts multi-scale features from both the RGB and high-frequency noise branches, followed by cross-fusion in both spatial and channel dimensions, making full use of the features extracted by the two branches.

We evaluate the model on several standard datasets, demonstrating its strong robustness and generalization capability.

2. Related Work

DeepFake Detection. Forged faces pose a significant th-reat to social security, making face forgery detection crucial. To address the unique characteristics of deepfake fa-ces, researchers proposed various detection methods. For instance, [36] used steganalysis and leverages spatial and inter-frame correlations to determine whether a face has been tampered with. Other methods focus on specific facial representations to detect authenticity, such as head pose, blinking, and mouth movements [26]. Employs frequency domain-aware decomposition, using frequency domain statistics to reveal deepfake artifacts in the frequency domain for detecting forged faces.

Generalization of Detection Enhanced by Synthetic Data. Although most existing methods perform well in detecting known manipulations, some studies found that these methods fail to generalize to faces forged by unknown techniques. To enhance the generalization capability of detectors, researchers adopted various strategies. For example, some researchers combined auxiliary localization tasks to guide the network to focus more precisely on the forged regions. In [12], one branch processes RGB input, while another branch uses DCT to extract high-frequency features from different frequency bands. The outputs of both branches are fused together to form more generalizable forgery features. Many approaches also integrated both RGB and frequency domain features. However, the high-fr-equency feature extraction methods mentioned above cannot adaptively fit the data to capture the most discriminative features. To improve detection generalization, different datasets of forged faces are needed. To this end, many synthetic datasets were created, such as FF++, FWA, CDF, DFD, DFDCP, and FFIW. Researchers used these datasets to train and cross-validate their methods, continuously improving the generalization of forgery detection.

Branch Feature Fusion. The branch fusion model aims to combine the analysis of image noise characteristics and manipulation traces to enhance the accuracy of image forgery detection [31, 40]. The model's two branches handle different aspects of the image: the first branch inputs the RGB source face and the second branch focuses on capturing high-frequency noise information. This noise exhibits unique patterns depending on the device and source, serving as an inherent feature of the image. The second branch inputs noise extracted by the SRM, concentrating on inconsistencies after image manipulation, analyzing how the manipulation disrupts the spatial characteristics of the noise and uncovering residual manipulation traces.

High-frequency features. The role of high-frequency features in image authenticity detection received widespread attention in recent years, especially in adversarial generative face forgery (such as DeepFake) technology. Existing studies shown that the distribution of real images and forged images in the high-frequency domain is significantly different [5], which provides key clues for detection algorithms. Early work mainly used frequency domain analysis methods such as discrete Fourier transform (DFT) and discrete cosine transform (DCT) to extract high-frequency features [7], but they were mostly limited to feature extraction in a single

domain. Recent studies pointed out that forged faces have inherent defects in the correlation of high-frequency features across domains [11], especially in high-frequency details such as edge sharpness and micro-texture consistency, which are prone to phase misalignment and energy distribution anomalies.

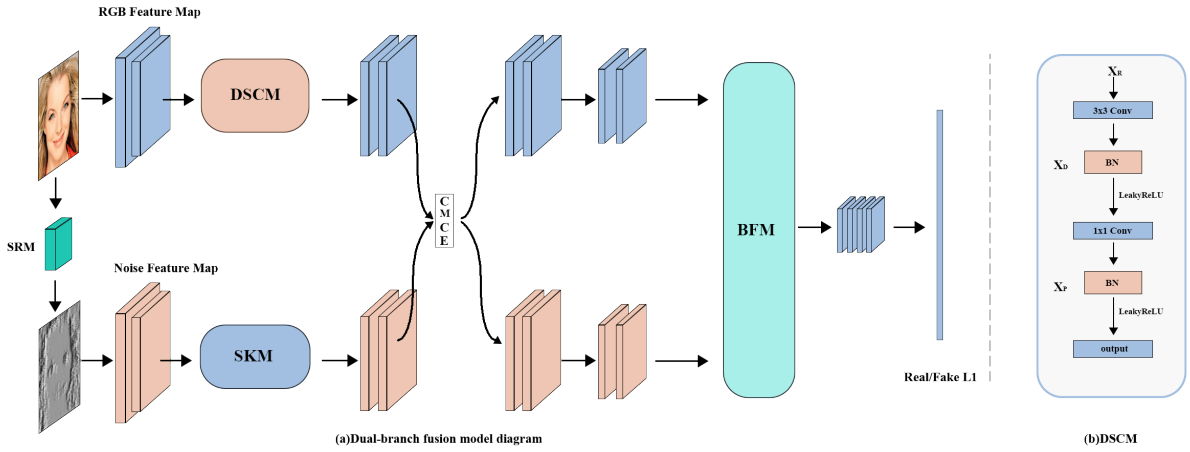

FIGURE 1. The diagram illustrates the overall flow of the model, which is mainly divided into two branches: one consisting of the RGB branch and the other consisting of high-frequency noise obtained through SRM processing. These two branches undergo CMCE cross-learning, and are finally fused through BFM.

3. METHOD

Figure 1 displays our proposed Enhanced Dual-Branch Fusion Model (EDFM). The model mainly consists of two branches: one is the RGB branch, and the other is the high-frequency noise branch obtained through SRM processing.

These two branches undergo cross-learning via Cross-Modality Consistency Enhancemen (CMCE) [31] and are finally fused through BFM for classification. In the shallow layers of the RGB branch, we take full advantage of the DSCM, which has the benefit of a small number of parameters and high performance, reducing computational overhead while ensuring stable performance for the RGB branch. Since SRM works through three fixed filters, it l-imits its ability to adapt to features. To address this, our method introduces the SKM, which allows each neuron in the convolutional neural network to adjust its receptive field size adaptively based on multi-scale input information, utilizing the dynamic selection mechanism of the neural network. This effectively resolves the issue of SRM's limit-ed adaptability. Thus, by using the SKM and combining it with SRM high-frequency noise, we further enhance SRM's ability to extract effective representations. Additionally, we leverage CMCE for interactive learning in the middle layers. Finally, we design a multi-scale channel-spatial attention fusion module to support the fusion of multi-scale channel-spatial attention between the RGB and high-frequency noise branches, fully integrating the representations learned from both branches for classification. In the following sections, we will discusseach component in detail.

3.1. **Depthwise separable convolutional module.** In the RGB branch, using the DSCM instead of regular convolution significantly improves the model's computational efficiency. DSCM breaks down regular convolution into two steps: depthwise convolution and pointwise convolution. This decomposition-greatly reduces the computational complexity, making the model more lightweight and efficient during training and fusion. Moreover, since each depthwise convolution operates only on a single channel and the pointwise convolution handles the fusion of information between channels, this design leads to a substantial reduction in the required parameters, which in turn lowers the model's storage and memory usage. In addition, while maintaining stable model accuracy, the feature extraction process of the RGB branch is highly efficient. Therefore, by using the DSCM in the RGB branch, the model can maintain its performance while greatly improving computational efficiency.

The specific implementation details of the module are shown in Figure 1(b). It contains a 2D convolution, normalization and LeakyReLU activation layer. First, we input the feature map X_R into the deep convolution layer to get X_D, Then, X_D is input into a point-by-point convolution layer to get the final output X_P.

3.2. **Selective kernel module in high-frequency noise branch.** In the dual-branch architecture, the SRM branch performs well in extracting noise features. However, during filtering for preprocessing, the use of three fixed convolution kernels limits the filter's ability to adaptively update to fit the data, which restricts its receptive field size for noise features. Therefore, we propose to use the SKM approach, which implements a dynamic selection mechanism in convolutional neural networks, allowing each neuron to adaptively adjust its receptive fields based on the input. This effectively addresses the shortcomings of the SRM in extracting noise features, enhancing its ability to capture noise across different receptive field sizes, and subsequently providing more effective features for classification.

SKM utilizes a building block called the selective kernel unit, which consists of multiple branches with different kernel sizes. These branches are fused through a softmax attention mechanism, guided by the information from these branches. This fusion process enables neurons to adaptively adjust their effective receptive field size based on the input. The specific structure of the module is displayed in Figure 2. First, the input feature X_s is processed by con-

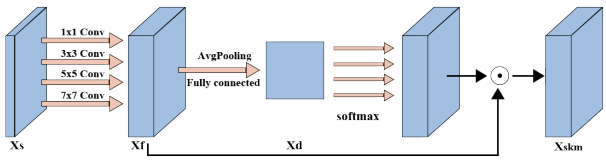

FIGURE 2. Selective Kernel Module. It adjusts the receptive field by dynamically selecting branches with different kernel sizes and attention mechanisms.

volution kernels of different sizes, where the convolution kernel sizes are 1×1, 3×3, 5×5, and 7×7, respectively, to extract feature maps of different receptive fields. The selection of 1×1, 3×3, 5×5, and 7×7 convolutional kernels in multi-scale feature extraction is driven by empirical and experimental considerations: 1×1 kernels enable lightweight channel-wise fusion, 3×3 kernels efficiently capture localized details while maintaining computational economy, 5×5 kernels model mid-range contextual patterns, and 7×7 kernels capture global dependencies. These scales collectively capture the typical spectrum of object scales in natural images,

spanning from fine-grained details to holistic structures. This multi-scale design ultimately achieves an optimal balance between computational efficiency and feature representation capacity. Then the outputs of the four different convolution kernels are summed to obtain the fused feature map, which is described as $X_f = \sum_{k=1}^{4} F_{2k-1}(X_s)$, where F_{2k-1} represents the convolution operation with $2k-1$ as the convolution kernel

Subsequently, global average pooling is applied to the fused feature map X_f, followed by a fully connected layer to obtain the dimension-reduced X_d. In addition, we need to compute the attention weights corresponding to each convolution kernel, apply each attention weight to the corresponding feature map, and then perform a weighted summation of all the feature maps to obtain the final output X_{skm}.

3.3. **Multi-scale branch fusion module.** Since this paper adopts an RGB and SRM dual-branch learning mode, effectively merging the features from both branches is a crucial issue. Traditional methods of simple addition or multiplication for feature fusion have significant limitations when handling these tasks, they are easily affected by noise and struggle to dynamically adjust the importance of features. We propose a multi-scale branch fusion module that combines multi-scale feature extraction with channel attention mechanisms to enhance the robustness of feature fusion. In this process, multi-scale convolutions can capture feature information at different scales, while the attention mechanism dynamically adjusts the importance of features, ensuring that key features receive higher weights during fusion.

Specifically, for the intermediate layer features T_1 and T_2 obtained from the RGB and SRM dual branches, both are convolved by using three different kernel sizes: 3×3, 5×5, and 7×7. After convolution, the results of the three convolutions are summed to produce a multi-scale feature map $T_1^{'}$

$$T_1^{'} = Conv3(T_i) + Conv5(T_i) + Conv7(T_i), i \in \{1,2\},$$

where $i=1$ and $i=2$ represent the intermediate feature maps of RGB and SRM.

After multi-scale feature extraction, pooling operations in the channel and spatial dimensions are also required. The cat operation in Figure 3 represents the connection of the channel-channel dimension and the space-space dimension of the two branches by using the cat operation in torch. Specifically, the obtained features $T_1^{'}$ and $T_2^{'}$ undergo global average pooling (AvgPool) and global max pooling (MaxPool) operations in both the channel and spatial dimensions, described as follows:

$$\begin{aligned} T_1^A &= AdvAvgPool\left(T_1^{'}\right), \\ T_1^M &= AdvMaxPool\left(T_1^{'}\right), \end{aligned} \tag{3.1}$$

where *AdvAvgPool* is the adaptive average pooling and *AdvMaxPool* is the adaptive maximum pooling, which can be adaptively changed to the specified HxW size output. The implementation of adaptive pooling is as follows:

$$S_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i,j).$$

$T_2^{'}$ performs the same operation (3.1) to obtain T_2^A and T_2^M.

$$W_1 = Soft\left(2DConv1\left(Cat\left(T_1^A, T_2^A\right)\right)\right) + Soft\left(1DConv1\left(Cat\left(T_1^M, T_2^M\right)\right)\right) \tag{3.2}$$

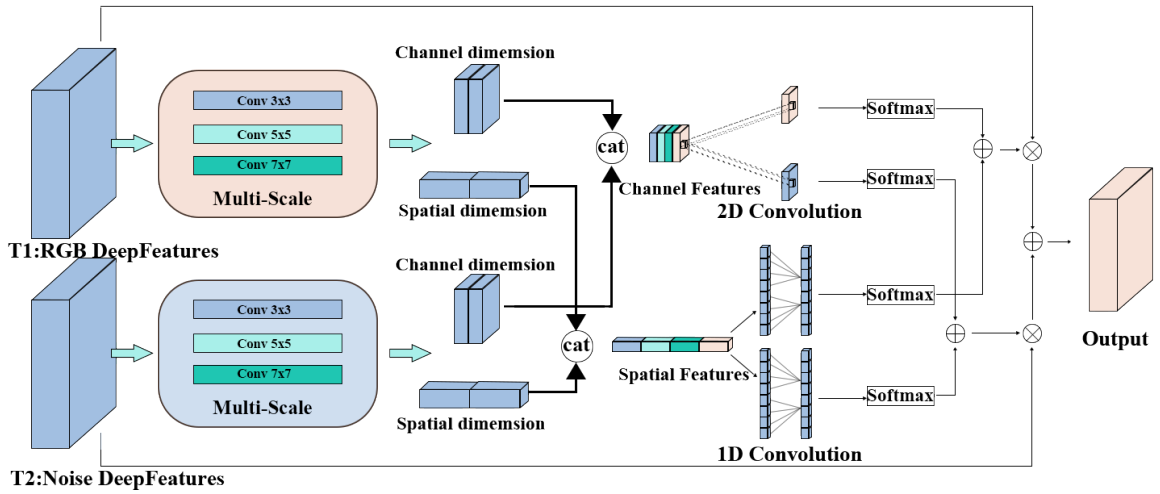

FIGURE 3. Diagram of the BFM via multi-scale convolution together with channel and spatial attention mechanisms.

and

$$W_2 = Soft\left(2DConv2\left(Cat\left(T_1^A, T_2^A\right)\right)\right) + Soft\left(1DConv2\left(Cat\left(T_1^M, T_2^M\right)\right)\right). \tag{3.3}$$

Among them, Cat represents the concatenate operation, 2DConv1(or 2) and 1DConv1(or 2) represent the convolution operation at time 1(or 2), and Soft represents SoftMax.

As shown in Figure 3, the results of channel pooling and spatial pooling calculated above are concatenated to obtain channel features and spatial features, respectively. Then, a one-dimensional convolution is applied to the concatenated spatial features to obtain two different spatial weights. These weights are normalized using softmax, ensuring that their sum is 1, which facilitates subsequent weighted fusion operations. Finally, using (3.2), (3.3), calculated channel weight W_1 and spatial weight W_2, the weighted addition and weighted multiplication of the two features are performed to obtain the fused feature map F. The specific description is $F = T_1W_1 + T_2W_2$.

3.4. **Loss function.** For the loss function, we use cross entropy loss to supervise network learning as follows: $L_C = -\left[y \log \hat{y} + (1-y) \log \hat{y}\right]$, where y is a binary label.

4. EXPERIMENT

4.1. **Implementation Settings.** we evaluate our model on five common forgery datasets: FaceForensics++ (FF++), Celeb-DF [16], DeepFake Detection [27] (DFD), and the Deepfake Detection Challenge [9] (DFDC). The experiments used the high-quality version (c23) of the FF++ dataset, which contains 4,000 forged videos generated by four algorithms: DeepFake (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT).

Implementation details: In the data preprocessing stage, we align the officially labeled FF++ dataset with the original videos and crop the facial regions. All experimental face images are crop to 299×299 and uniformly normalized to [0, 1]. We utilize common augmentations such as flipping, contrast adjustment, scaling, and blurring. Additionally, to increase dataset diversity while ensuring annotations are aligned with images, we employ random cropping. We

used Xception [28] for pre-training model initialization, with the Adam [41] optimizer set to betas 0.9 and 0.999, and epsilon 1e-8. The batch size is set to 32, with a learning rate of 0.0002. All experiments are implemented using PyTorch on the NVIDIA RTX 4090 24GB platform.

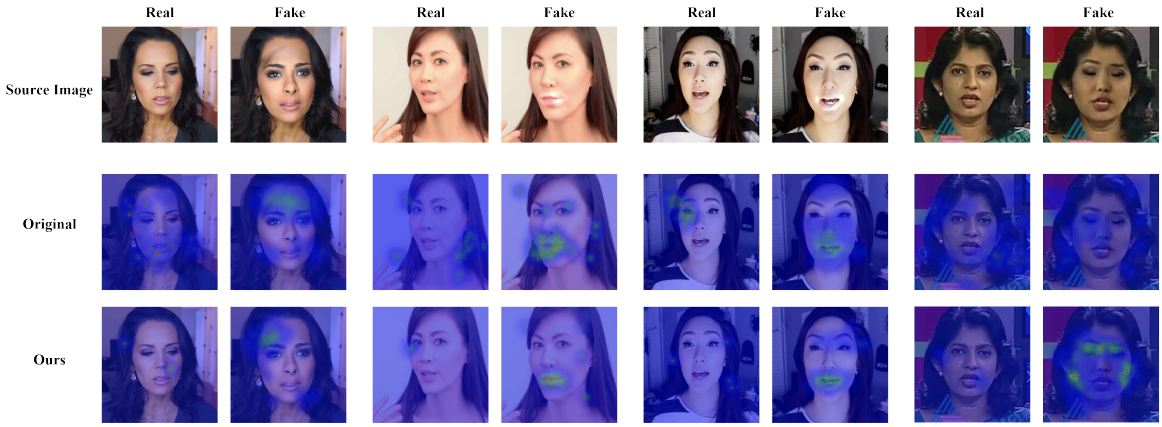

FIGURE 4. Comparison of the differences in the heatmaps before and after the improvement. Ours refers to the EDFM, and Original refers to the dual-branch basic model of direct additive fusion.

4.2. **Evaluation. Intra-dataset Evaluation**. Table 1 lists the results of testing four different methods on the four forgery techniques in the FF++ dataset. It can be seen that when the traing set and the testing set are the same dataset, our method can achieve high results comparable to other methods, and our method achieves performance comparable to other methods with similarly high results when test within the same dataset.

Cross-dataset evaluation. When applied to cross-dataset scenarios, reveals that some methods experience significantly low AUC (Area Under the Curve) during testing. For example, in Table 1, the DCL method shows an AUC of 52.1% when test from NT to F2F, while Xception displays an AUC of only 49% in the DF to FS scenario. Similarly, Face X-ray achieves only 45.8% AUC when evaluate from FS to DF . In contrast, our method demonstrates balanced performance, with AUC values consistently above 70%, indicating strong stability.

Additionally, when evaluating DF to other datasets and NT to other datasets, our method generally achieves the best AUC results. For instance, when evaluating NT to the other four forgery methods, the average AUC reaches an impressive 94.7%. Notably, in the DF to F2F cross-dataset scenario, our method shows a significant improvement, with an AUC that is 9.4% higher than the second-best DCL method. Furthermore, As shown in Table 2, our method achieves excellent results in AUC test. As seen in Figure 4, the enhanced model more accurately captures the forged regions, providing more valuable feature information for detection. Heatmaps are generated by extracting feature maps from the last convolutional layer of a deep learning model, then combining them with the target category weights from the fully connected layer through weighted summation to create activation maps. These activation maps undergo activation processing, normalization, and upsampling before being superimposed onto the original image. This visualizes the model's decision-critical regions (highlighted areas), intuitively revealing the key areas the model focuses on when making classifications.

TABLE 1. AUC evaluation across datasets on FF++(HQ).The boldface indicates the best result, and the underlined one indicates the suboptimal result.

Training Set	Model	Testing Set(AUC)				
		DF	F2F	FS	NT	Avg
DF	Xception [28]	0.993	0.736	0.490	0.736	0.554
	Face X-ray [13]	0.987	0.764	0.600	0.698	0.762
	DCL [32]	**1.00**	0.771	0.610	0.782	0.790
	Ours	0.998	**0.865**	**0.680**	**0.857**	**0.850**
F2F	Xception [28]	0.803	0.994	0.762	0.696	0.813
	Face X-ray [13]	0.630	0.984	**0.938**	**0.945**	**0.874**
	DCL [32]	**0.919**	0.992	0.596	0.667	0.793
	Ours	0.844	**0.996**	0.804	0.786	0.857
FS	Xception [28]	0.664	0.888	0.994	0.713	0.814
	Face X-ray [13]	0.458	**0.961**	0.981	**0.957**	**0.839**
	DCL [32]	**0.741**	0.698	0.995	0.526	0.740
	Ours	0.682	0.891	**0.997**	0.729	0.824
NT	Xception [28]	0.799	0.813	0.731	0.991	0.834
	Face X-ray [13]	0.705	0.917	0.910	0.925	0.864
	DCL [32]	0.912	0.521	0.783	0.990	0.801
	Ours	**0.922**	**0.955**	**0.918**	**0.994**	**0.947**

Cross-method evaluation. We compare our method with LAE, ClassNSeg, and ForensicTrans, as shown in Table 3. First, all methods are train on NT and then test on F2F and NT. It can be observed that our method achieves a 13% higher performance on F2F compared to the second best, ForensicTrans, further proving the effectiveness of our method in cross-method scenarios on the FF++ dataset. Even in within-method testing, our method outperforms ForensicTrans by 4%.

These results are achieved through our BFM module, which preserves RGB-extracted information during fusion, enabling high accuracy within methods. Additionally, the SKM module accurately retains high-frequency noise features, enhancing the model's generalization ability, leading to the 13% improvement. Therefore, the experimental results further confirm the superior generalization capability of our method.

TABLE 2. The FF++ training and testing datasets are evaluated with other methods, and the evaluation metric is AUC.

Training Set	Model	Testing Set(AUC)				
		DF	F2F	FS	NT	Avg
FF++	Xception [28]	0.994	0.995	0.994	0.995	0.994
	Face X-ray [13]	0.991	0.993	0.992	0.993	0.992
	DCL [32]	**1.00**	0.990	0.999	0.976	0.991
	SOLA [8]	**1.00**	0.995	**1.00**	**0.998**	**0.998**
	SBIs [33]	**1.00**	0.999	0.999	0.988	0.996
	Ours	0.998	**0.999**	0.995	0.993	0.996

TABLE 3. Comparison of Acc evaluation with LAE, ClassNSeg, and ForensicTrans in NT to F2F, in High Quality (HQ) case.

Training Set	Model	Testing Set(Acc)	
		F2F(HQ)	NT(HQ)
NT(HQ)	LAE [20]	0.72	0.87
	ClassNSeg [22]	0.65	0.88
	ForensicTrans [12]	0.74	0.92
	Ours	**0.87**	**0.96**

Cross-dataset comparison.The cross-dataset comparison results are shown in Table 5. We train the model on FF++ and then test it on four cross-datasets: DFD, DFDC, CelebDF, and DF1. 0, comparing the results with Xception and Face X-ray. This test is more valuable since the four datasets have less similarity to FF++, requiring better generalization from the model to achieve good results.

The experimental results in Table 4 indicate that the model demonstrates strong robustness against image compression at various quality levels. On original images, t-he model achieves high accuracy and AUC, with 84.5% and 94%, respectively, showing excellent detection performance. When the compression quality is reduced to 90% and 70%, the model's performance slightly decreases but still maintains good accuracy (81.7% and 77.4%) and AUC (89% and 85%), indicating strong resistance to mild and moderate compression. However, at a compression quality of 50%, the accuracy and AUC drop significantly to 70.2% and 81%, respectively, suggesting that heavy compression impacts the model's performance due to the loss of image details. Overall, the model retains high detection capability across different compression conditions, demonstrating good noise resistance and robustness.

As seen from the results in Table 5, our method demonstrates high stability, achieving the best average AUC score of 76.2% across the four datasets. Unlike using Xception on CelebDF, our method avoids significant drop and consistently maintaining good results. This cross-dataset comparison confirms that our method improves generalization in some scenarios. For example, in the DFDC dataset, our method outperforms Face X-ray by 4.5%. It also maintains strong stability, with balanced results across all four datasets, highlighting the robustness of the model.

Finally, we compare our method's generalization performance on the cross-dataset CD2 with recent models like F3Net, FWA, MADD, and MTD-Net. As shown in Table 6, our approach achieves better results than recent models like F3Net and FWA, further confirming that the improved dual-branch fusion and SKM modules enhance the model's generalization capability.

TABLE 4. Performance evaluation of the model on the FF++ dataset at different JPEG compression qualities

Compression quality	Acc	AUC	PSNR	SSIM
Original Image	0.845	0.94	–	–
90%	0.817	0.89	39.8	0.98
70%	0.774	0.85	36.5	0.95
50%	0.702	0.81	31.2	0.90

TABLE 5. Cross-dataset evaluation from FF++ to other dataset.The boldface indicates the best result, and the underlined one indicates the suboptimal result.

Training Set	Model	Testing Set(AUC)				
		DFD	DFDC	CelebDF	DF1.0	AVG
FF++	Xception [28]	0.831	0.679	0.594	0.689	0.698
	Face X-ray [13]	**0.856**	0.700	0.742	<u>0.723</u>	0.755
	Luo [18]	0.839	<u>0.732</u>	<u>0.810</u>	0.690	<u>0.767</u>
	Ours	<u>0.842</u>	**0.745**	**0.842**	**0.732**	**0.790**

5. ABLATION EXPERIMENT

To evaluate the effectiveness of the three modules, we conduct ablation experiments. As shown in Table 7, we first train the FF++ dataset using the RGB, SRM, and dual-branch models

to establish baseline data. Then, we test the same training with the DSCM, SKM, and BFM modules in succession. The process is as follows: we first evaluate the individual RGB and SRM branches separately, then evaluate the fusion results of the two branches. As seen from the third row in Table 7, the fusion results on DF and NT outperform the single-branch approach.

Starting from the fourth row, we test the methods proposed in this paper: DSCM, SKM, and BFM. From the analysis of the table 7, it can be concluded that the basic DSCM module shows certain improvements compared to the regular dual-branch model, especially with a 4.7% increase in the NT evaluation. Next, the combination of the DSCM and SKM modules further improves performance compared to the single DSCM module, with a 1.4% increase in both FS and NT evaluations. This confirms that the SKM module enhances SRM's ability to extract high-frequency features. Finally, with the addition of the BFM module, the overall is improved even more, particularly in the NT evaluation, where it increases from 95.5% to 96.7%. By analyzing the results from the last three rows, it is clear that with the addition of each module, the overall model performance steadily improves, demonstrating the effectiveness of each module.

The primary rationale for selecting Leaky ReLU as the activation function lies in its superior performance, achieving the highest AUC (0.955) on the test set compared to other activation functions. Additionally, Leaky ReLU addresses the "dying neuron" issue inherent to standard ReLU by preserving small gradients in negative input regions, while simultaneously mitigating the vanishing gradient problem commonly observed in Sigmoid and Tanh functions. This dual mechanism enhances training effectiveness in deep neural networks. Its enhanced nonlinear expressive capability enables the model to capture more sophisticated feature representations, thereby ultimately improving detection performance in face forgery identification tasks.

As shown in Table 9.From the experimental data, it is evident that the multi-scale convolutional kernel combination of 1×1, 3×3, 5×5, and 7×7 achieves the best performance in forgery detection. This combination yields the highest AUC values across all test sets, with 0.922 on DF, 0.955 on P2F, 0.918 on FS, and 0.994 on NT, significantly outperforming other configurations. In contrast, removing certain kernel sizes (e.g., excluding 5×5 or 7×7) leads to performance degradation, indicating that multi-scale features are crucial for forgery detection. The 1×1, 3×3, 5×5, and 7×7 combination effectively captures multi-scale features, thereby enhancing detection performance.

TABLE 6. By training on FF++ (c23), we cross the dataset to CD2 and compare with the nearest method(F3Net,FWA,MADD,MTD-Net,Dual-branch,GFF). The evaluation result is AUC.

Model	Training Set	Testing Set(AUC)
		CD2
F3Net [26]	FF++(c23)	65.17
FWA [17]	Self-made	57.32
MADD [39]	FF++(c23)	67.44
MTD-Net [38]	FF++(c23)	70.12
Dual-branch [21]	FF++(c23)	73.41
GFF [8]	FF++(c23)	65.20
Luo [18]	FF++(c23)	66.23
Shuai [31]	FF++(c23)	67.38
Ours	FF++(c23)	68.48

TABLE 7. Ablation study on FF++. The metric is AUC. Results in grey indicate performance within the dataset.

Method	DF	F2F	FS	NT
RGB	0.803	0.994	0.762	0.696
SRM	0.758	0.994	0.913	0.858
Dual-branch	0.805	0.994	0.910	0.894
DSCM	0.821	0.995	0.931	0.941
DSCM + SKM	0.833	0.996	0.945	0.955
DSCM+ SKM + BFM	**0.838**	**0.996**	**0.956**	**0.967**

TABLE 8. The test results of the model under different activation functions.

Training Set	Activation Function	Testing Set(AUC)
		FS
NT	Tanh	0.892
	ReLU	0.920
	Sigmoid	0.885
	Leaky ReLU	**0.955**

TABLE 9. Detection results across datasets at different scales.

Training set	Size combination	Testing set(AUC)			
		DF	F2F	FS	NT
NT	3x3,5x5,7x7	0.884	0.899	0.843	0.921
	1x1,5x5,7x7	0.897	0.928	0.884	0.953
	1x1,3x3,7x7	0.901	0.848	0.855	0.909
	1x1,3x3,5x5	0.912	0.937	0.874	0.945
	1x1,3x3,5x5,7x7	**0.922**	**0.955**	**0.918**	**0.994**

6. CONCLUSION

We proposed a novel dual-branch fusion model that enhances the fusion of the RGB and SRM branches using multiple modules. This approach significantly improves generalization and balance in cross-dataset face deepfake detection. We use DSCM to extract enhanced features of RGB images and then combine it with SKM to obtain information that is more conducive to detecting high-frequency noise. Finally, we proposed to perform multi-scale feature extraction on the branches and then perform channel and spatial fusion module BFM to achieve the effect of enhancing robustness. We achieved stable performance and enhanced the model's overall generalization ability across datasets. Finally, comprehensive experiments demonstrate the effectiveness of each module, and our method outperforms other detection methods.

Acknowledgments
This work was partially supported by China-Croatia Bilateral Science & Technology Cooperation Project.

REFERENCES

[1] Z. Akhtar, T.L. Pendyala, V.S. Athmakuri, Video and audio deepfake datasets and open issues in deepfake technology: Being ahead of the curve, Forensic Sciences, 4 (2024),

289-377.

[2] A. Busacca, M.A. Monaca, Deepfake: Creation, Purpose, Risks. In: Marino, D., Monaca, M.A. (eds) Innovations and Economic and Social Changes due to Artificial Intelligence: The State of the Art. Studies in Systems, Decision and Control, vol 222. Springer, Cham, 2023.

[3] T. Buckley, B. Ghosh, V. Pakrashi, A feature extraction & selection benchmark for structural health monitoring, Structural Health Monitoring, 22 (2023), 2082-2127.

[4] W. Chen, K. Shi, Multi-scale attention convolutional neural network for time series classification, Neural Networks, 136 (2021), 126-140.

[5] R. Durall, M. Keuper, F.J. Pfreundt, J. Keuper, Unmasking deepfakes with simple features, arXiv preprint arXiv:1911.00686, 2019.

[6] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Transactions on Information Forensics and Security, 7 (2012), 868-882.

[7] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, Leveraging frequency analysis for deep fake image recognition, In: Proceedings of the 37th International Conference on Machine Learning, PMLR 119 (2020), 3247-3258.

[8] J. Fei, Y. Dai, P. Yu, T. Shen, Z. Xia, J. Weng, Learning second order local anomaly for general face forgery detection, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2022.

[9] L. Guarnera, O. Giudice, F. Guarnera, et al., The face deepfake detection challenge, Journal of Imaging, 8 (2022), 263.

[10] A. Heidari, N.J. Navimipour, H. Dag, M. Unal, Deepfake detection using deep learning methods: A systematic and comprehensive review, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14 (2024), e1520.

[11] Y. He, N. Yu, M. Keuper, M. Fritz, Beyond the spectrum: Detecting deepfakes via re-synthesis, arXiv preprint arXiv:2105.14376, 2021.

[12] J. Li, H. Xie, J. Li, Z. Wang, Y. Zhang, Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6458-6467, 2021.

[13] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face x-ray for more general face forgery detection, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001-5010, 2020.

[14] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510-519, 2019.

[15] Y. Li, M.C. Chang, S. Lyu, In ictu oculi: Exposing AI generated gake face videos by detecting eye blinking, arXiv preprint arXiv:1806.02877, 2018.

[16] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df (v2): A new dataset for deepfake forensics, arXiv:1909.12962, 2019.

[17] Y. Li, Exposing deepfake videos by detecting face warping artif acts, arXiv preprint arXiv:1811.00656, 2018.

[18] Y. Luo, Y. Zhang, J. Yan, W. Liu, Generalizing face forgery detection with high-frequency features, In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 16317-16326, 2021.

[19] X. Liu, S. Wang, S. Lu, Z. Yin, X. Li, L. Yin, J. Tian, W. Zheng, Adapting feature selection algorithms for the classification of chinese texts, Systems, 11 (2023), 483.
[20] D. Maignan, L. Yuening, Towards generalizable forgery detection with localityaware autoencoder, arXiv preprint arXiv:1909.05999, 2019.
[21] I. Masi, A. Killekar, R.M. Mascarenhas, S.P. Gurudatt, W. AbdAlmageed, Two-branch recurrent network for isolating deepfakes in videos, In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12352. Springer, Cham, 2020.
[22] H.H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos, In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1-8, IEEE, 2019.
[23] R. Natsume, T. Yatagawa, S. Morishima, RSGAN: Face Swapping and editing using face and hair representation in latent spaces, arXiv preprint arXiv:1804.03447, 2018.
[24] L.A Passos, D. Jodas, K.A. Costa, L.A. Souza Júnior, D. Rodrigues, J. Del Ser, D. Camacho, J.P. Papa, A review of deep Learning-based approaches for deepfake content detection, Expert Systems, 41 (2024), e13570.
[25] W. Qin, T. Lu, L. Zhang, S. Peng, D. Wan, Multi-Branch deepfake detection algorithm based on Fine-Grained features, Computers, Materials & Continua, 77 (2023), 467-490.
[26] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, In: European Conference on Computer Vision, pp. 86-103. Springer, 2020.
[27] M.S. Rana, M.N. Nobi, B. Murali, A.H. Sung, Deepfake detection: A systematic literature review, IEEE Access, 10 (2022), 25494-25513.
[28] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1-11, 2019.
[29] A.D. Samuel-Okon, O.I. Akinola, O.O. Olaniyi, O.O. Olateju, S.A. Ajayi, Assessing the effectiveness of network security tools in mitigating the impact of deepfakes AI on public trust in media, Archives of Current Research International, 24 (2024), 355-375.
[30] J. Stehouwer, H. Dang, F. Liu, X. Liu, A. Jain, On the detection of digital face manipulation, arXiv:1910.01717, 2019.
[31] C. Shuai, J. Zhong, S. Wu, F. Lin, Z. Wang, Z. Ba, Z. Liu, L. Cavallaro, K. Ren, Locate and verify: A two-stream network for improved deepfake detection, In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 7131-7142, 2023.
[32] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, R. Ji, Dual contrastive learning for general face forgery detection, In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2316-2324, 2022.
[33] K. Shiohara, T. Yamasaki, Detecting deepfakes with self-blended images, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18720-18729, 2022.
[34] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, Information Fusion, 64 (2020), 131-148.

[35] L. Verdoliva, Media forensics and deepfakes: an overview, IEEE Iournal of Selected Topics in Signal Processing, 14 (2020), 910-932.

[36] X. Wu, Z. Xie, Y. Gao, Y. Xiao, SSTNet: Detecting manipulated faces through spatial, steganalysis and temporal features, In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2952-2956, IEEE, 2020.

[37] C. Yang, S.N. Lim, One-shot domain adaptation for face generation, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5921-5930, 2020.

[38] J. Yang, A. Li, S. Xiao, W. Lu, X. Gao, Mtd-net: Learning to detect deepfakes images by multi-scale texture difference, IEEE Transactions on Information Forensics and Security, 16 (2021), 4234-4245.

[39] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, Multi-attentional deepfake detection, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2185-2194, 2021.

[40] D. Zhang, W. Zhu, X. Ding, G. Yang, F. Li, Z. Deng, Y. Song, SRTNet: a spatial and residual based two-stream neural network for deepfakes detection, Multimedia Tools and Applications, 82 (2023), 14859-14877.

[41] T. Zhou, W. Wang, Z. Liang, J. Shen, Face forensics in the wild, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5778-5788, 2021.