A Dual-Branch CNN for Robust Detection of AI-Generated Facial Forgeries

1st Xin Zhang
Computer Science Department
University of Southern Maine
Portland, United States
xin.zhang@maine.edu

2nd Yuqi Song
Computer Science Department
University of Southern Maine
Portland, United States
yuqi.song@maine.edu

3rd Fei Zuo
The Department of Computer Science
University of Central Oklahoma
Edmond, United States
fzuo@uco.edu

***Abstract*—The rapid advancement of generative AI has enabled the creation of highly realistic forged facial images, posing significant threats to AI security, digital media integrity, and public trust. Face forgery techniques—ranging from face swapping and attribute editing to powerful diffusion-based image synthesis—are increasingly being used for malicious purposes such as misinformation, identity fraud, and defamation. This growing challenge underscores the urgent need for robust and generalizable face forgery detection methods as a critical component of AI security infrastructure. In this work, we propose a novel dual-branch convolutional neural network for face forgery detection that leverages complementary cues from both spatial and frequency domains. The RGB branch captures semantic information, while the frequency branch focuses on high-frequency artifacts that are difficult for generative models to suppress. A channel attention module is introduced to adaptively fuse these heterogeneous features, highlighting the most informative channels for forgery discrimination. To guide the network's learning process, we design a unified loss function—FSC Loss—that combines focal loss, supervised contrastive loss, and a frequency center margin loss to enhance class separability and robustness. We evaluate our model on the DiFF benchmark, which includes forged images generated from four representative methods: text-to-image, image-to-image, face swap, and face edit. Our method achieves strong performance across all categories and outperforms average human accuracy. These results demonstrate the model's effectiveness and its potential contribution to safeguarding AI ecosystems against visual forgery attacks.**

***Index Terms*—AI Security, Multimedia Forensics, Computer Vision, Face Forgery Detection, Deepfake Detection**

I. Introduction

The rapid advancement of generative artificial intelligence (AI), particularly in deep generative models such as GANs and diffusion models, has transformed the way digital images are created [1]. These models can synthesize highly realistic, photorealistic images that are virtually indistinguishable from real ones, even to trained observers [2]. While these technologies have led to remarkable progress in creative industries, advertising, entertainment, and virtual reality, they also pose significant societal threats when misused for malicious purposes [3]. Image forgeries can be exploited for identity theft, political propaganda, harassment, and large-scale misinformation campaigns, eroding trust in digital media and compromising personal and national security [4]. With the widespread dissemination of forged content on social platforms, the development of **robust image forgery detection systems** is more critical than ever.

As illustrated in the DiFF benchmark [5], modern image forgeries can be generated using four primary methods: (a) text-to-image (T2I) generation from textual prompts, (b) image-to-image (I2I) synthesis via fine-tuned diffusion models, (c) face swapping (FS) where the identity in a target image is replaced with a source face, and (d) face editing (FE) where attributes such as age, expression, or gender are altered using prompt-based modifications. All these techniques leverage advanced generative models that produce forgeries with highly natural textures, making detection extremely challenging. The detailed generation pipelines are shown in Fig. 1.

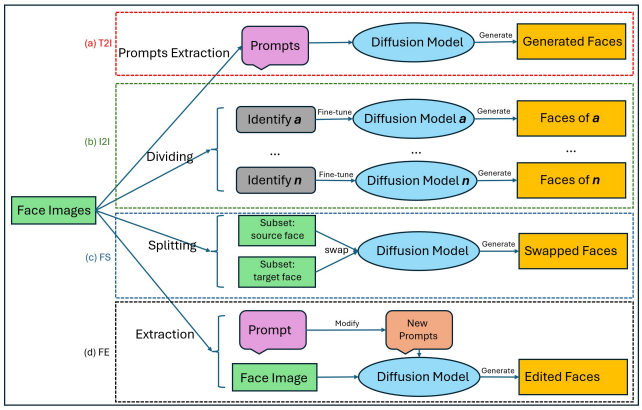

Fig. 1. Detailed pipelines for generating diffusion-based facial forgeries under four conditions. (a) T2I: Prompts are extracted from face images and fed into a diffusion model to synthesize new faces. (b) I2I: The dataset is divided by identities (e.g., identity a to n), and each identity is used to fine-tune a separate diffusion model to generate diverse images of that specific identity. (c) FS: The dataset is split into source and target subsets; the diffusion model swaps identity features between the two subsets to produce realistic swapped faces. (d) FE: Prompts and face images are extracted, the prompts are modified, and the updated prompts are passed to a diffusion model to edit attributes such as expression, age, or style while preserving the core facial features.

The detection challenge is evident in the DiFF dataset, where human performance is only 59.65% for text-to-image, 59.63% for image-to-image, 51.50% for face swapping, and 45.53% for face editing [5]. Such results show that even human observers struggle to reliably identify diffusion-based forgeries, highlighting the need for automated detection sys-

tems capable of capturing subtle artifacts in both spatial and frequency domains.

In this paper, we propose a dual-branch CNN-based forgery detection model that integrates spatial and frequency information using channel attention fusion. The RGB branch learns semantic and structural features, while the frequency branch focuses on high-frequency residuals where imperceptible generative artifacts are often present. The two branches are fused via a channel attention module, which adaptively assigns weights to the most informative features from each branch. This design enables the network to effectively leverage complementary cues for forgery detection. To further enhance discriminative feature learning, we introduce FSC Loss (Focal–Supervised Contrastive–Center Loss), which consists of three components: focal loss [6], supervised contrastive loss [7], and frequency center margin loss [8]. The combination is designed to not only emphasize hard-to-classify samples (via focal loss) but also to improve the compactness and separability of feature embeddings (via supervised contrastive loss) while ensuring that frequency-domain features of real and fake images remain well-distinguished (via frequency center margin loss).

Our main contributions are as follows:

- We propose a dual-branch ResNet-based architecture with channel attention fusion, effectively combining spatial and frequency cues for robust forgery detection.
- We design a novel FSC Loss, integrating focal, supervised contrastive, and frequency center margin objectives to improve classification performance and feature separability.
- We conduct extensive experiments on the DiFF dataset, demonstrating that our approach outperforms existing detectors across all four forgery categories (T2I, I2I, FS, FE).

II. Realted Work

The research on face forgery detection has evolved from early handcrafted forensic features to deep learning-based methods that can capture subtle artifacts introduced by modern generative models. Traditional techniques—such as splicing detection, copy-move detection, and noise inconsistency analysis—relied on pixel-level anomalies like unnatural transitions or lighting inconsistencies [9], [10]. While effective for conventional image tampering, these methods fail on AI-generated forgeries, which exhibit smooth textures and high photorealism.

With the rise of GAN-based deepfakes, CNNs became the foundation of forgery detection. XceptionNet demonstrated strong performance by capturing pixel-level manipulation patterns [11], [12], while Face X-ray [13] improved detection by focusing on blending artifacts near manipulated boundaries. To better generalize to unseen forgeries, frequency-domain methods have been explored. It has been shown that GAN-generated images exhibit distinctive anomalies in the Fourier spectrum [14], [15]. F3-Net [16] combines spatial and frequency cues in a multi-branch architecture, leading to robust performance across manipulation types.

More recently, attention mechanisms [17] and transformer-based models [18] have been applied to focus on manipulation-prone regions and long-range dependencies. However, diffusion models pose a new challenge, as their iterative denoising process produces fewer frequency artifacts and more natural textures [19], [20], making traditional GAN-trained detectors less effective. This motivates hybrid approaches which captures both spatial structures and high-frequency residual cues to address diffusion-based forgeries.

III. Proposed Method

The overall pipeline of our proposed face forgery detection model is illustrated in Figure 2. Our framework adopts a dual-branch architecture that leverages both RGB and frequency-domain cues to enhance detection robustness.

First, the input face image is simultaneously processed through two pathways:

- The RGB branch receives the original image and extracts spatial features using a ResNet-50 backbone (with the classification head removed).
- The frequency branch first converts the RGB image to the frequency domain using a Fast Fourier Transform (FFT). The resulting spectrum is then passed through a ResNet-34 backbone to capture complementary frequency-based features.

The outputs of these two branches are concatenated along the channel dimension to form a unified representation. To adaptively emphasize the most informative channels, we apply a Channel Attention Module to the concatenated features. This module learns to reweight channels based on their relevance to the forgery detection task.

The attention-enhanced feature is then passed through fully connected layers to perform binary classification, predicting whether the input image is real or fake.

A. Dual-branch Design

To enhance the model's ability to detect subtle and diverse artifacts introduced by modern generative techniques, we adopt a dual-branch architecture that processes input images through both RGB and frequency domains. This design leverages complementary information: spatial texture patterns in the RGB space and spectral artifacts in the frequency domain, resulting in improved robustness and generalization.

Motivation for Dual-Branch Design. Purely spatial (RGB-based) detectors can struggle to identify forgeries generated by high-fidelity generative models such as diffusion-based text-to-image or face-editing systems. These forgeries often exhibit highly natural visual appearances, with little to no low-level anomalies observable in the pixel domain. However, as highlighted by prior works such as [16], [21], generative models often leave behind frequency inconsistencies due to non-linear activations, interpolation, and upsampling operations. These spectral anomalies are hard for human observers to detect but can be exploited algorithmically.

By integrating both the RGB and frequency views, the model captures:

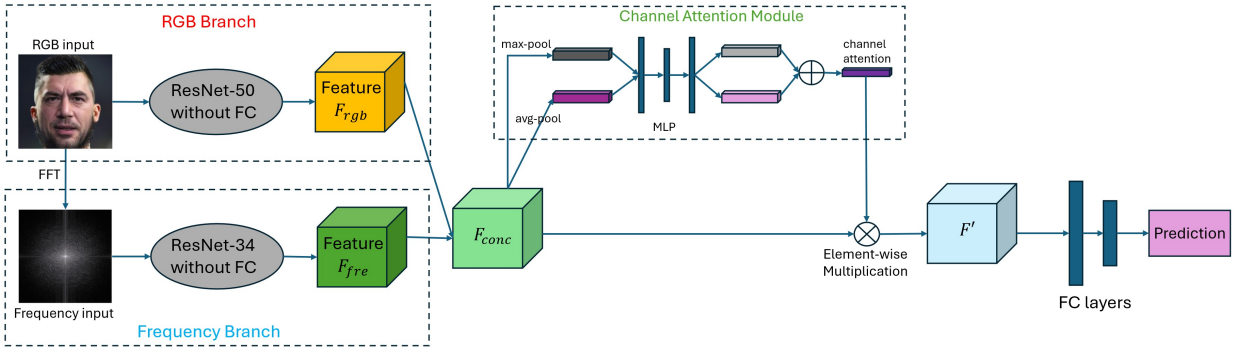

Fig. 2. **Overview of the proposed dual-branch detection framework.** The RGB image is processed through a ResNet-50 backbone, while its frequency representation (obtained via FFT) is passed through a ResNet-34. The resulting features are concatenated and refined using a Channel Attention Module, which adaptively emphasizes informative channels. The fused feature is then pooled and passed through fully connected layers for real-vs-fake prediction.

- Local texture patterns and spatial dependencies via the RGB branch.
- Global periodic and spectral characteristics via the frequency branch.

This dual design allows the model to detect a wider range of forgery traces and improves generalizability across manipulation types.

Frequency Input Generation. To capture subtle frequency artifacts introduced by generative models, we transform each RGB image into a frequency-domain representation. This is achieved through the following steps:

- The RGB image $I_{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$ is first converted into a single-channel grayscale image $I_{\text{gray}} \in \mathbb{R}^{H\times W}$ using the luminance-preserving weighted sum:

$$I_{\text{gray}} = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \tag{1}$$

  where R, G, and B represent the red, green, and blue channels of the input image, respectively. These coefficients are derived from the ITU-R BT.601 standard and reflect the human eye's varying sensitivity to different wavelengths.
- We then compute the magnitude spectrum of the grayscale image using the Fast Fourier Transform (FFT) [22]. To enhance stability and suppress extreme values, we apply logarithmic scaling:

$$I_{\text{fre}} = \log\left(1 + |\mathcal{F}(I_{\text{gray}})|\right) \tag{2}$$

  Here, $\mathcal{F}(\cdot)$ denotes the 2D FFT operator, and $|\cdot|$ represents the magnitude of the complex frequency coefficients. The resulting $I_{\text{fre}} \in \mathbb{R}^{H\times W}$ is used as the input to the frequency branch of our model.

Feature Representation and Fusion. Given an RGB input image $I \in \mathbb{R}^{3\times H\times W}$, our model extracts two complementary types of features through a dual-branch architecture:

- **RGB Branch:** The input is processed through a ResNet-50 backbone (excluding the final fully connected layer), resulting in a spatial feature map:

$$F_{\text{rgb}} \in \mathbb{R}^{2048\times 7\times 7}$$

- **Frequency Branch:** The precomputed frequency representation I_{fre} is passed through a ResNet-34 backbone to produce:

$$F_{\text{fre}} \in \mathbb{R}^{512\times 7\times 7}$$

The two feature maps are concatenated along the channel dimension and modulated by a channel attention map to highlight informative features. Specifically, we compute:

$$F' = M_c \odot \text{Concat}(F_{\text{rgb}}, F_{\text{fre}}), \quad M_c \in \mathbb{R}^{2560\times 1\times 1}$$

where $\odot$ stands for the element-wise multiplication, and $F' \in \mathbb{R}^{2560\times 7\times 7}$ is the attention-weighted feature map. Global average pooling (GAP) is then applied to obtain the final feature vector:

$$f = \text{GAP}(F') \in \mathbb{R}^{2560}$$

which is passed through fully connected layers to produce a binary output indicating whether the image is real or forged.

To compute the channel attention map M_c, we follow the strategy of the Convolutional Block Attention Module (CBAM) [23]. Given the concatenated feature map F_{conc}, we first apply both average pooling and max pooling along the spatial dimensions to generate two descriptors:

$$f_{\text{avg}} = \text{AvgPool}(F_{\text{conc}}), \quad f_{\text{max}} = \text{MaxPool}(F_{\text{conc}})$$

Both f_{avg} and $f_{\text{max}} \in \mathbb{R}^{2560\times 1\times 1}$ are passed through a multilayer perceptron (MLP) consisting of two fully connected layers with ReLU activation. The MLP outputs are then summed and passed through a sigmoid function to obtain the final channel attention map:

$$M_c = \sigma(\text{MLP}(f_{\text{avg}}) + \text{MLP}(f_{\text{max}}))$$

This attention mechanism adaptively emphasizes the most informative channels for the classification task.

B. The FSC loss function

To effectively supervise the training process and encourage the network to learn discriminative features across both RGB and frequency domains, we propose a novel loss function called **FSC Loss**, which integrates three components: Focal

Loss [6], Supervised Contrastive Loss [7], and Frequency Center Margin Loss [8]. The overall loss function is defined as:

$$\mathcal{L}_{\text{FSC}} = \mathcal{L}_{\text{focal}} + \lambda_1 \cdot \mathcal{L}_{\text{supcon}} + \lambda_2 \cdot \mathcal{L}_{\text{f-center}}$$

where λ_1 and λ_2 are weighting coefficients that balance the contribution of each term.

- **Focal Loss** ($\mathcal{L}_{\text{focal}}$). This term addresses class imbalance and emphasizes hard-to-classify samples. It modifies the standard cross-entropy loss with a focusing parameter γ and a weighting factor α. For a binary classification problem with label $y \in \{0, 1\}$ and predicted probability p, the focal loss is:

$$\mathcal{L}_{\text{focal}} = \begin{cases} -\alpha(1-p)^{\gamma}\log(p), & \text{if } y = 1 \\ -(1-\alpha)p^{\gamma}\log(1-p), & \text{if } y = 0 \end{cases}$$

- **Supervised Contrastive Loss** ($\mathcal{L}_{\text{supcon}}$). This loss improves feature representation by encouraging embeddings from the same class to be close together, and those from different classes to be far apart. Given a batch of samples, for each anchor feature z_i, we consider the set of positive features $P(i)$ (same class) and contrast them against all other features $A(i)$ (excluding i itself). The loss is:

$$\mathcal{L}_{\text{supcon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}$$

  Here, $\cdot$ denotes dot product similarity, τ is a temperature parameter controlling the sharpness of the distribution, and z_i is treated as the anchor feature. This helps create compact clusters for each class in the embedding space.
- **Frequency Center Margin Loss** ($\mathcal{L}_{\text{f-center}}$). This term improves class separation in the frequency feature space. It consists of two parts: (1) pulling frequency features f_i closer to their corresponding class centers c_{y_i}, and (2) pushing the centers of different classes away from each other by at least a fixed margin m. The loss is formulated as:

$$\mathcal{L}_{\text{f-center}} = \sum_{i=1}^{N} \|f_i - c_{y_i}\|_2^2 + \mu \sum_{j \neq k} \max(0, m - \|c_j - c_k\|_2)^2$$

  The first term ensures that features of the same class are compact, while the second term penalizes class centers that are too close. This helps the model distinguish between real and fake images based on frequency-domain cues.

By combining these three complementary components, the FSC loss enables the model to learn robust and semantically meaningful features from both spatial and spectral perspectives.

IV. Experiments

A. Dataset

We evaluate our method using the DiFF benchmark [5], a recently introduced dataset dedicated to facial image forgery detection. DiFF is designed to reflect real-world forgery generation techniques by incorporating four representative manipulation categories: Text-to-Image (T2I), Image-to-Image (I2I), Face Swap (FS), and Face Edit (FE). These categories span a wide range of generative pipelines used in practice.

- **Text-to-Image (T2I):** Images are synthesized from prompts using pretrained diffusion models such as FreeDoM [24].
- **Image-to-Image (I2I):** Images are generated from real faces via fine-tuned diffusion models like LoRA [25], preserving general structure but altering identity or expression.
- **Face Swap (FS):** Identity information from one face is transplanted onto another using methods like DiffFace [26].
- **Face Edit (FE):** Attribute-level changes (e.g., age, expression, gender) are applied using prompt-based image editing techniques such as CoDiff [27] and Imagic [28].

For each category, DiFF includes both pristine (real) and manipulated (fake) samples in a 1:1 ratio. The dataset is curated to avoid identity leakage between training and test sets by following an identity-disjoint protocol. Evaluation splits are provided to support both in-domain and cross-domain benchmarking, enabling robust assessment of model generalization across forgery types.

We select the DiFF benchmark for its diversity, realism, and coverage of modern forgery techniques.Unlike earlier datasets that focus only on GAN-based manipulations, DiFF comprehensively evaluates detection performance across diffusion-generated and edited content, better reflecting today's forgery landscape.

B. Experimental Results

In-domain Performance. In-domain performance refers to the setting where the model is trained and evaluated on the same type of forgery, allowing it to specialize in learning manipulation-specific features. This evaluation serves as a fundamental benchmark to assess the model's upper-bound performance and its ability to capture intrinsic patterns within each forgery category, prior to testing its generalization to unseen domains. As summarized in Table I, our proposed method achieves strong in-domain detection results across all forgery types, performing comparably to, and in some cases outperforming, state-of-the-art baselines. These results demonstrate that the dual-branch architecture, combined with the FSC loss, effectively leverages both spatial and frequency-domain cues to identify forgeries with high accuracy when trained and tested within the same manipulation domain. This highlights the model's ability to learn highly discriminative representations under in-distribution conditions.

TABLE I
IN-DOMAIN DETECTION ACCURACY (%) ON THE DIFF BENCHMARK. EACH MODEL IS TRAINED AND EVALUATED ON THE SAME FORGERY TYPE USING AN 80/20 TRAIN-TEST SPLIT. THE BEST-PERFORMING METHOD IS HIGHLIGHTED IN BOLD.

Method	T2I-T2I	I2I-I2I	FS-FS	FE-FE
Xception [11]	93.32	98.92	99.95	**99.95**
F^3-Net [16]	99.60	99.50	**99.98**	99.60
EfficientNet [29]	99.89	99.80	99.87	99.24
DIRE [30]	95.04	99.88	99.09	99.87
Ours	**99.91**	**99.89**	99.94	99.07

Cross-domain Performance. Cross-domain evaluation assesses the model's ability to generalize across different types of forgery generation methods. Specifically, models are trained on one subset (e.g., T2I) and tested on other subsets (e.g., I2I, FS, FE). This setting reflects real-world scenarios where the type of forgery encountered during deployment may differ from that seen during training. As shown in Table. II, our model exhibits strong generalization ability across different forgery generation methods. When trained on the T2I subset, it achieves the best performance on all other test sets: 90.21% on I2I, 47.98% on FS, and 73.70% on FE. Under the I2I training setting, our model outperforms baselines on T2I (91.17%), FS (47.91% vs. 41.51% for DIRE), and FE (50.21% vs. 46.19% for F^3-Net). When trained on FS, it again achieves the best performance across T2I (42.30%), I2I (39.21%), and FE (36.88%), substantially surpassing the strongest baseline DIRE. Lastly, under the FE training condition, our method reaches the highest score on T2I (85.32%), and performs competitively on I2I (78.34% vs. 79.12%) and FS (68.97% vs. 70.81%). These results consistently demonstrate our model's robustness in cross-domain scenarios, where training and testing are conducted on different manipulation techniques.

Comparison with Human Performance. Table III presents a comparison between human performance and our model across four forgery generation categories. The second to fifth columns correspond to detection performance on each type of forgery: T2I (text-to-image), I2I (image-to-image), FS (face swap), and FE (face edit). Human performance values are taken from the DiFF benchmark [5], which conducted a large-scale user study involving 70 participants. Each participant was instructed to classify the authenticity of 200 randomly selected images with a 50:50 split between real and fake images, ensuring no identity repetition to avoid bias. In total, 14,000 human classification decisions were collected. We report the average accuracy per forgery category. Our model's results are computed by averaging the detection accuracies for each category across **all corresponding cross-domain test settings**. The comparison shows that our method consistently outperforms human observers in all categories. These results demonstrate the effectiveness of our model in detecting highly realistic forgeries that are often difficult even for humans to identify.

Ablation Study. To evaluate the effectiveness of individual components in our architecture, we conduct ablation experiments under three configurations:

- w/o *Fre-Branch*: Removes the frequency branch, using only the RGB stream for feature extraction.
- w/o $\mathcal{L}_{\text{f-center}}$: Excludes the frequency center margin loss from the training objective.
- w/o M_c: Eliminates the channel attention mechanism; features from both branches are directly concatenated without attention weighting.

As shown in Table. IV, removing the frequency branch leads to the most significant drop in performance, with an average decrease of 13.4% across test sets, demonstrating the crucial role of frequency information. Eliminating $\mathcal{L}_{\text{f-center}}$ causes noticeable drops as well, particularly on the FE set (6.79%↓), confirming that frequency-specific supervision strengthens discriminability. Omitting M_c results in degraded performance on all subsets, with a 3.1% drop on average, indicating the importance of attention-guided feature fusion. These results highlight that each component (frequency representation, frequency-aware loss, and attention fusion) contributes meaningfully to the model's overall performance.

TABLE II
CROSS-DOMAIN GENERALIZATION RESULTS. EACH MODEL IS TRAINED ON ONE SUBSET AND EVALUATED ON ALL SUBSETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST RESULTS ARE UNDERLINED.

Methods	Train Subset	Testing Set			
		T2I	I2I	FS	FE
Xception	T2I	-	86.85	34.65	23.28
F^3-Net	T2I	-	88.50	<u>45.07</u>	<u>71.06</u>
EfficientNet	T2I	-	<u>89.72</u>	21.49	49.63
DIRE	T2I	-	84.07	35.15	50.86
Ours	T2I	-	**90.21**	**47.98**	**73.70**
Xception	I2I	<u>87.82</u>	-	36.82	33.39
F^3-Net	I2I	87.23	-	40.62	<u>46.19</u>
EfficientNet	I2I	84.39	-	19.47	27.46
DIRE	I2I	86.20	-	<u>41.51</u>	42.01
Ours	I2I	**91.17**	-	**47.91**	**50.21**
Xception	FS	23.17	24.47	-	10.17
F^3-Net	FS	<u>35.43</u>	30.39	-	20.79
EfficientNet	FS	16.88	22.17	-	10.21
DIRE	FS	16.08	<u>36.27</u>	-	<u>32.68</u>
Ours	FS	**42.30**	**39.21**	-	**36.88**
Xception	FE	80.84	**79.12**	**70.81**	-
F^3-Net	FE	<u>82.32</u>	76.92	56.27	-
EfficientNet	FE	80.41	63.06	66.62	-
DIRE	FE	56.70	59.22	43.78	-
Ours	FE	**85.32**	<u>78.34</u>	<u>68.97</u>	-

TABLE III
COMPARISON WITH HUMAN PERFORMANCE ACROSS DIFFERENT FORGERY TYPES.

Method	T2I	I2I	FS	FE
Human	59.65	59.63	51.50	45.53
Ours	**72.92**	**69.25**	**54.95**	**53.60**

V. CONCLUSION

With the rapid evolution of generative models, forged facial images have become increasingly realistic and widespread, posing serious threats to digital media integrity, public trust,

TABLE IV
ABLATION STUDY OF KEY COMPONENTS. THIS TABLE QUANTIFIES THE IMPACT OF REMOVING INDIVIDUAL COMPONENTS OF OUR MODEL: THE FREQUENCY BRANCH, THE FREQUENCY CENTER LOSS ($\mathcal{L}_{\text{F-CENTER}}$), AND THE CHANNEL ATTENTION MODULE (M_c). THE PERFORMANCE DROP ACROSS SUBSETS HIGHLIGHTS THE EFFECTIVENESS OF EACH DESIGN.

Configurations	Testing Set			
	T2I	I2I	FS	FE
w/o **Fre-Branch**	94.23	82.31	33.68	42.78
w/o $\mathcal{L}_{\text{f-center}}$	99.06	88.17	43.01	66.91
w/o M_c	97.45	88.29	45.88	67.74
Ours	**99.91**	**90.21**	**47.98**	**73.70**

and security. Detecting such forgeries—especially those generated by advanced diffusion-based methods—is essential for safeguarding the authenticity of visual content in the AI era.

To address this challenge, we propose a novel detection framework capable of handling a wide range of forgery generation techniques, including text-to-image synthesis, image-to-image translation, face swapping, and face editing. Our method leverages a dual-branch CNN architecture that captures complementary cues from both RGB and frequency domains, fused via a channel attention mechanism to enhance informative features. Additionally, we introduce FSC Loss, a composite objective combining focal loss, supervised contrastive loss, and frequency center margin loss, to promote discriminative and robust feature learning.

Comprehensive experiments on the DiFF benchmark demonstrate the effectiveness of our approach. The model consistently performs well across both in-domain and cross-domain settings, and significantly outperforms average human detection accuracy. These results highlight the importance and feasibility of building generalized and trustworthy forgery detection systems for real-world applications.

REFERENCES

[1] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE journal of selected topics in signal processing*, vol. 14, no. 5, pp. 910–932, 2020.

[2] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[3] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[4] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[5] H. Cheng, Y. Guo, T. Wang, L. Nie, and M. Kankanhalli, "Diffusion facial forgery detection," in *Proceedings of the 32nd ACM international conference on multimedia*, 2024, pp. 5939–5948.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[7] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[8] J. Li, H. Xie, J. Li, Z. Wang, and 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*, 2021, pp. 6458–6467.

[9] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016, pp. 5–10.

[10] H. Farid, "Exposing digital forgeries from jpeg ghosts," *IEEE transactions on information forensics and security*, vol. 4, no. 1, pp. 154–160, 2009.

[11] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[12] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.

[13] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[14] T. Dzanic, K. Shah, and F. Witherden, "Fourier spectrum discrepancies in deep network generated images," *Advances in neural information processing systems*, vol. 33, pp. 3022–3032, 2020.

[15] S. Hu, Y. Li, and S. Lyu, "Exposing gan-generated faces using inconsistent corneal specular highlights," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2500–2504.

[16] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*. Springer, 2020, pp. 86–103.

[17] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[18] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2tr: Multi-modal multi-scale transformers for deepfake detection," in *Proceedings of the 2022 international conference on multimedia retrieval*, 2022, pp. 615–623.

[19] S. Pashine, S. Mandiya, P. Gupta, and R. Sheikh, "Deep fake detection: survey of facial manipulation detection solutions," *arXiv preprint arXiv:2106.12605*, 2021.

[20] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[21] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*. PMLR, 2020, pp. 3247–3258.

[22] E. O. Brigham, *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988.

[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[24] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang, "Freedom: Training-free energy-guided conditional diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 174–23 184.

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[26] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, and K. Lee, "Diffface: Diffusion-based face swapping with facial guidance," *Pattern Recognition*, vol. 163, p. 111451, 2025.

[27] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6080–6090.

[28] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6007–6017.

[29] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[30] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 445–22 455.