

D²Fusion: Dual-domain Fusion with Feature Superposition for Deepfake Detection

Xueqi Qiu^a, Xingyu Miao^{#a}, Fan Wan^a, Haoran Duan^b, Tejal Shah^b,
Varun Ojha^{b,*}, Yang Long^{a,*}, Rajiv Ranjan^b

^a*Department of Computer Science, Durham University, UK.*

^b*School of Computing, Newcastle University, UK.*

Abstract

Deepfake detection is crucial for curbing the harm it causes to society. However, current Deepfake detection methods fail to thoroughly explore artifact information across different domains due to insufficient intrinsic interactions. These interactions refer to the fusion and coordination after feature extraction processes across different domains, which are crucial for recognizing complex forgery clues. Focusing on more generalized Deepfake detection, in this work, we introduce a novel bi-directional attention module to capture the local positional information of artifact clues from the spatial domain. This enables accurate artifact localization, thus addressing the coarse processing with artifact features. To further address the limitation that the proposed bi-directional attention module may not well capture global subtle forgery information in the artifact feature (e.g., textures or edges), we

*Corresponding author # equal contribution

Email addresses: xueqi.qiu@durham.ac.uk (Xueqi Qiu),
xingyu.miao@durham.ac.uk (Xingyu Miao[#]), fan.wan@durham.ac.uk (Fan Wan),
haoranduan28@gmail.com (Haoran Duan), tejal.shah@newcastle.ac.uk (Tejal Shah),
varun.ojha@newcastle.ac.uk (Varun Ojha), yang.long@durham.ac.uk (Yang Long),
raj.ranjan@newcastle.ac.uk (Rajiv Ranjan)

employ a fine-grained frequency attention module in the frequency domain. By doing so, we can obtain high-frequency information in the fine-grained features, which contains the global and subtle forgery information. Although these features from the diverse domains can be effectively and independently improved, fusing them directly does not effectively improve the detection performance. Therefore, we propose a feature superposition strategy that complements information from spatial and frequency domains. This strategy turns the feature components into the form of wave-like tokens, which are updated based on their phase, such that the distinctions between authentic and artifact features can be amplified. Our method demonstrates significant improvements over state-of-the-art (SOTA) methods on five public Deepfake datasets in capturing abnormalities across different manipulated operations and real-life. Specifically, in intra-dataset evaluations, D²Fusion surpasses the baseline accuracy by nearly 2.5%. In cross-manipulation evaluations, it exceeds the baseline AUC by up to 6.15%. In multi-source manipulation evaluations, it exceeds the SOTA methods by up to 14.62% in P-value, 10.26% in F1-score and 15.13% in R-value. In cross-dataset experiments, it exceeds the baseline AUC by up to 6.25%. For potential applications, D²Fusion can help improve content moderation on social media and aid forensic investigations by accurately identifying the tampered content.

Keywords: Attention mechanism, Deepfake detection, Dual-domain fusion, Feature superposition

1. Introduction

Deepfake is an emerging facial video forgery technique which is used for creating fake videos based on AI technology. The videos generated by Deepfake can show people saying and doing things that never actually happened, and it is difficult for the human eye to distinguish the authenticity of these videos. Although Deepfake technology could be employed for productive endeavours such as film production, art style transformation, and virtual reality in the education field [1], Deepfake is often maliciously used for activities such as financial fraud, pornographic revenge, fake political news, and other activities that lead to property loss, damage the reputation of celebrities and misguide public opinion [2]. Therefore, research around Deepfake detection is critical for cybersecurity, law, and politics, as well as for individual and social well-being.

Various Deepfake detection methods rely on a vanilla binary classifier to extract artifact features for detection [3, 4, 5, 6, 7]. This approach to coarse processing artifact features causes the network to predominantly focus on low-level manipulated trajectories [8], resulting in diminished precision in pinpointing the forged areas. Similarly, frequency information methods [9, 10, 11] may ignore high-frequency signals, leading to the inability to capture some subtle forgery hints. In terms of combining different domain forgery traces, dual-branch network [12], dynamic graph [13], and feature-disentangled representation learning [14] are employed. However, the lack of sufficient generalization in processed features implies that while these methods may excel in detecting operation-specific artifacts, they often show a marked decline in performance when faced with unseen forgery techniques.

In this work, we present the D²Fusion framework, an innovative approach aimed at generalized facial forgery detection. To enhance artifact features, we introduce a bi-directional attention module. This module utilizes average pooling in both vertical and horizontal directions to extract local positional information from spatial domain, enabling precise localization of local artifacts. However, this attention module struggles to capture global details like texture and edge information. To address this, we integrate a fine-grained spectral attention module, employing discrete cosine transform (DCT) in multi-spectral partitioning. This approach retains high-frequency information including texture and edge information, thereby enhancing artifact feature details. To further strengthen the fusion of this complementary information from diverse domains, we propose a strategy of feature superposition. This strategy iteratively aggregates feature components according to their positional information. It amplifies the distinctiveness between the artifact feature and the authentic feature, thus making the features more generalized to various face manipulation algorithms and real-world scenarios. In summary, the contributions of this work are multi-fold:

- We investigate forged trajectories in artifact features from the feature-level perspective, and propose a bi-directional attention module that captures local information in features.
- To obtain global detailed information in artifact features, we utilize multi-spectral components with DCT and propose a fine-grained frequency attention module.
- To fuse information from different domains, we propose feature components in the form of wave-like tokens and update these tokens based on

their phase, amplifying the difference between authentic and artifact features.

- We show that our method outperforms state-of-the-art methods on five public datasets. Extensive experimental results demonstrate that our Deepfake detection model can capture abnormalities more effectively.

2. Related Works

2.1. Deepfake Generation

Deepfake transformations can be categorized into two categories: face swapping and face reenactment. Face swapping can replace the face in the source image with the same face shape and features as the target face. Face reenactment is a face synthesis task in which the facial expressions and postures of the target face are transferred to the source face while preserving the appearance and details of the source face [15].

In the early stage of Deepfake, based on computer graphics approaches [16, 17, 18], the forgery methods exhibited significant drawbacks, like the loss of facial expressions and unnatural appearance [19]. After that, deep-learning based techniques are gaining popularity in generating synthetic media. The most common architecture of existing Deepfake generation [20, 21] is an auto-encoder with two separate decoders for facial identity exchange [22]. Recent Deepfake research also takes advantage of Generative Adversarial Networks (GAN) [23, 24, 25] for improved synthesis authenticity.

2.2. Deepfake Detection

Early Deepfake detection relied on visual artifacts, the images generated by GANs or Autoencoders must be further distorted to align the original

Table 1: A comparison of recent research methods.

Method	Year	Category	Feature Type	Feature Pre-processing	Strength	Weakness
Distorted Boundary[26]	2018	Visual artifacts	Spatial	✗	Utilize facial area distortion	Low detection accuracy and limited to basic visual artifacts
Face X-ray [6]	2020			✗	Reveal the blending boundary	
ID-Reveal [27]	2021	Object detection	Spatial	✗	Use prior biometric characteristics of a depicted identity	Over-reliance on object features and not as effective as deep learning based detection methods
ICT [28]	2022			✗	Combine transformer with identity feature extraction	
Object Representations [29]	2023			✗	Look for object level coherence in spatial dimensions	
MAT [30]	2021	Deep learning	Spatial	Aggregate the different level features with the attention maps	Formulate deepfake detection as a fine-grained classification problem	Dependent on specific domain artifacts and vulnerability to adversarial attacks
SBI [7]	2022			✗	Effective detection of most basic face swapping operations	
FADE [31]	2023			Utilize multi-dependency graph	Can be integrated with some existing frame-level methods	
CADDM[32]	2023			✗	Address the implicit identity leakage issue	
GFF [33]	2021		Frequency	Residual guided spatial attention	Utilize image noises	
FreqGAN [11]	2022			✗	Not limited to the training settings	
E-TAD [34]	2024			✗	Focus on texture inconsistencies	
FreqNet [35]	2024			Frequency convolutional layer	The detector can consistently prioritize and focus on high-frequency information	
IDM [36]	2024	Spatio-frequency	Spatio-frequency	Feature recombination and residual calculation	Amplify the illumination inconsistency	Generalization ability still needs further improve.
SFDG [13]	2023			Dynamic graph learning	Discover the relationships with a graph-based relation-reasoning approach	
FDML [8]	2023			Feature-disentangling	Only forgery-relevant features are used	
TAN-GFD [37]	2023			Multi-level adaptive noise mining	Analyze and utilize texture and noise information	
LSDA [38]	2024			Distill knowledge	Enhance generalization by simulating forgery feature variations	

faces in the source video. Such a transformation leaves obvious visual artifact cues of distorted facial shape [26] and blending boundary [6]. Especially, [6] is foundational in exposing visual forgery traces. It achieves detection effectiveness by focusing on the presence of fusion boundaries, thereby reducing computational load. However, detections based on visual artifacts are limited to basic face-warping trajectories, they can not deal with more complex deepfake operations.

Another kind of method focuses on object detection by identifying inconsistencies between the subject and the background in the spatial dimension. [27] introduces the ID-Reveal, which leverages metric learning and adversarial

training strategies to learn the facial dynamic features of specific individuals. [28] introduces the Identity Consistency Transformer (ICT) and addresses the utilization of high-level semantic information by detecting identity inconsistencies between the inner and outer facial regions. [29] utilizes vision transformers to extract object representations and detect object-level spatial inconsistencies both intra-frame and inter-frame. Object detection based on inconsistencies between the background and the subject offers strong interpretability. Still, it may overfit detection toward fixed object features and be less effective than deep learning based detection methods.

Deep learning based detectors are currently the primary approaches [30, 39, 7, 31, 32], offering more complex feature extraction and classification compared to other detection approaches. In a single spatial domain, represented by MAT [30], which aggregates the different level features with the attention maps and first formulates Deepfake detection as a fine-grained classification problem. SBI [7] adopts EfficientNetB4 [5] as the classifier and offers a data augmentation strategy that uses self-blended images for training data. However, SBI typically generates full-face images, thus remaining confined to specific types of deepfake generation. Subsequent work [32] improved upon SBI by employing multi-scale face swapping to reduce reliance on identity information. Meanwhile, theoretical underpinnings for identifying forged images using frequency domain are established in [9, 10]. The work in [9] demonstrates that the mean amplitude across different frequency bands in genuine images differs from that in fake images, identifying frequency domain information as a valuable detection indicator. Further, the study in [10] extensively analyzes the frequency spectrum across various

GAN-based forgery techniques using Discrete Cosine Transform (DCT). Increasingly, subsequent works have leveraged frequency domain information for detection [11, 33, 34, 35, 36], among which [34] can focus on texture inconsistencies and [35] can consistently prioritize and focus on high-frequency information. However, the simplistic processing of single-domain information limits the detection capabilities of models and vulnerability to adversarial attacks.

Furthermore, a part of the works employs both the spatial and frequency information [13, 8, 37, 38], but their generalization ability still needs further improvement. [13] introduces the use of dynamic graphs to explore high-order relations between spatial and frequency features, which requires a great number of parameters to store the relationship between spatial and frequency domains. TAN-GFD [37] combines multi-scale texture difference features and regional noise inconsistency features, when performing cross-dataset detection, its performance declines. The model in [8] tries to automatically separate forgery-relevant features in the spatial domain and frequency domain with a feature-disentangling strategy but requires a priori knowledge as a guide. [38] introduces a Deepfake detection method based on Latent Space Data Augmentation (LSDA), which expands the forgery space by constructing and simulating variations of forgery features within the latent space but also increases the computational workload.

In general, the above methods still lead to the model relying on the specific manipulated patterns and scenarios, we prefer to have the model learn the more generalized differences between authentic and artifact features across different domains with their properties, thereby enabling the model to achieve

excellent detectability in different forgery operations and scenarios.

3. Method

3.1. Architecture

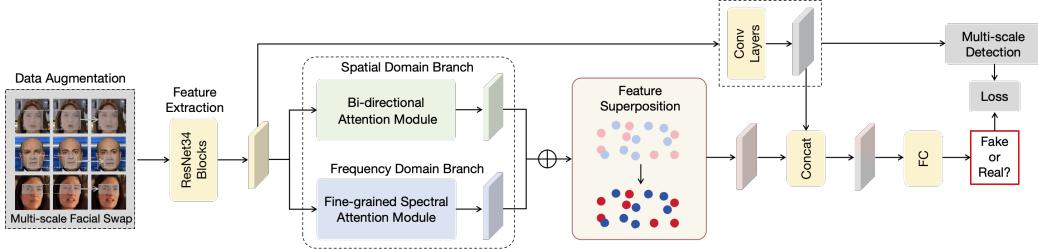


Figure 1: The pipeline of our D²Fusion.

In this section, we introduce the complete structure of D²Fusion (Figure 1). Before facial forgery detection, we employ the multi-scale facial swap module [32] for data augmentation. Subsequently, we utilize ResNet-34 as the backbone to extract facial features.

The extracted features are then handled by the Dual-domain Attention area, comprising the bi-directional attention module and the fine-grained spectral attention module. The bi-directional attention module innovatively encodes features along both horizontal and vertical directions, subsequently generating an intermediate feature map through convolutional processing that integrates spatial information from both directions, effectively preserves the spatial relationships in forged images and accurately localizes manipulated areas. This is coupled with a fine-grained spectral attention module that converts the features into the frequency domain. It splits the feature into multi-spectral components with Discrete Cosine Transform (DCT), enhancing the focus on high-frequency details containing global artifact information.

Rather than simply learning from combined features, our network employs a feature superposition strategy for feature processing. Within this strategy, the positional information of tokens is defined as their phase, and their actual values correspond to the amplitude. This configuration facilitates the iterative fusion of tokens based on phase and amplitude, significantly boosting the discriminability of features. Additionally, we employ a multi-scale detection model [32] as a shortcut to refine the detection outcomes.

3.2. Preliminary

We employ the multi-scale facial swap module [32] to produce new forgery images for data augmentation. This module employs a randomly sized sliding window that aims to target the region most likely to contain artifacts:

$$x_t, y_t = \underset{x,y}{\operatorname{argmax}} \sum_{i=x}^{x+hy+w} \sum_{j=y}^h \text{DSSIM}(I_f, I_s)_{i,j}, \quad (1)$$

where x_t, y_t represents the top-left position of the sliding window on the image, I_f is identified as the fake image, and I_s as the source image. $\text{DSSIM}(\cdot)$ [40, 41] serves to quantify the differences between two images by evaluating brightness l , contrast c , and structure s :

$$\begin{aligned} \text{DSSIM}(I_f, I_s)_{i,j} &= \frac{1}{1 - [l(I_f, I_s)_{i,j}]^\alpha [c(I_f, I_s)_{i,j}]^\beta [s(I_f, I_s)_{i,j}]^\gamma}, \\ l(I_f, I_s)_{i,j} &= \left(\frac{2\mu_{I_f}\mu_{I_s} + C_1}{\mu_{I_f}^2 + \mu_{I_s}^2 + C_1} \right)_{i,j}, \\ c(I_f, I_s)_{i,j} &= \left(\frac{2\phi_{I_f}\phi_{I_s} + C_2}{\phi_{I_f}^2 + \phi_{I_s}^2 + C_2} \right)_{i,j}, \\ s(I_f, I_s)_{i,j} &= \left(\frac{\rho_{I_f, I_s} + C_3}{\rho_{I_f}\rho_{I_s} + C_3} \right)_{i,j}, \end{aligned} \quad (2)$$

where μ is mean, ϕ is standard deviation, ρ is the covariant, and C_1, C_2, C_3 are all constants used to maintain l, c, s stability.

By omitting the region under the sliding window on the fake image, a mask M is computed, facilitating the data augmentation with a new fake image I'_f with blending method ξ as:

$$I'_f = \xi(I_f, I_s, M). \quad (3)$$

3.3. Dual-domain Attention

3.3.1. Bi-directional Attention Module

Previous work fails to consider the positional relationships between local facial parts, while we present the bi-directional attention (Figure 2) that preserves spatial relationships in forged images and accurately localizes manipulated regions.

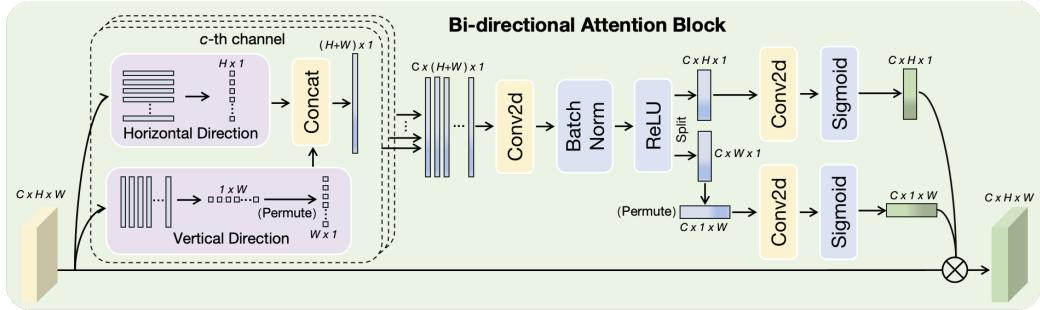


Figure 2: The structure of bi-directional attention block.

Following [42], given the input $X \in \mathbb{R}^{C \times H \times W}$, the module first encodes each channel along the horizontal and vertical directions, yielding a compact feature representation across different directions. For applying horizontal average pooling to c -th channel feature $x_c \in \mathbb{R}^{H \times W}$, the module uses a spatial extent pooling kernel shaped $(1, W)$. The output z_c^h of x_c at height h can be expressed as:

$$z_c^h = \frac{1}{W} \sum_{i=0}^{W-1} x_c^h, \quad (4)$$

and we obtain the transferred feature $Z_c^H = [z_c^1, z_c^2, \dots, z_c^H], Z_c^H \in \mathbb{R}^{H \times 1}$.

Similarly, for applying vertical average pooling to x_c , the module uses another spatial extent pooling kernel shaped $(H, 1)$. The output of the c -th channel feature x_c at width w can be transformed into:

$$y_c^w = \frac{1}{H} \sum_{i=0}^{H-1} x_c^i, \quad (5)$$

and we obtain the transferred feature $Y_c^W = [y_c^1, y_c^2, \dots, y_c^W], Y_c^W \in \mathbb{R}^{1 \times W}$.

We perform a permute operation on Y_c^W to make $Y_c^W \in \mathbb{R}^{W \times 1}$ for subsequent concatenation operation $\epsilon[\cdot, \cdot]$:

$$q_c = \epsilon[Z_c^H, Y_c^W], \quad (6)$$

$q_c \in \mathbb{R}^{(H+W) \times 1}$ is the concatenation feature at c -th channel, and we can obtain the total concatenation feature $q_n = [q_1, q_2, \dots, q_C], q_n \in \mathbb{R}^{C \times (H+W) \times 1}$. To enhance the expressive power of q_n , this module employ the 1×1 convolutional transformation function F_1 and ReLU function δ , to obtain the intermediate feature map f_c :

$$f_c = \delta(F_1(q_n)). \quad (7)$$

The transformations described facilitate the intermediate feature map f_c generation process, which encodes spatial information across both horizontal and vertical dimensions, ensuring stability and convergence.

To refocus each segment of the feature on a specific spatial direction, we split the feature map f_c along the spatial dimension, yielding two separate tensors: $f_c^h \in \mathbb{R}^{C \times H \times 1}$ and $f_c^w \in \mathbb{R}^{C \times W \times 1}$. Also, for subsequent multiplication operations, we permute $f_c^w \in \mathbb{R}^{C \times W \times 1}$ as $f_c^w \in \mathbb{R}^{C \times 1 \times W}$. Further

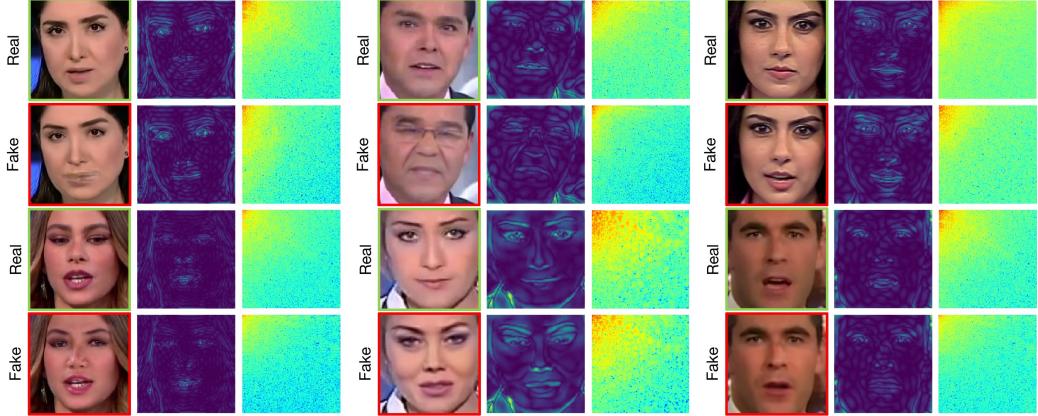


Figure 3: This image displays the high-frequency information from real and fake images, with the corresponding spectrum. Generated images contain strong high frequencies components (visible as the more blue region in the lower right corner), while real images contain more lower frequency components (visible as the more brightened region in the top left corner.)

expanding f_c^h and f_c^w using two more 1×1 convolutional transformations F_h and F_w , with the Sigmoid function σ , and the reweighted output X_{bi} from the bi-directional attention module is obtained as:

$$X_{bi} = X \times \sigma(F_h(f_c^h)) \times \sigma(F_w(f_c^w)). \quad (8)$$

3.3.2. Fine-grained Spectral Attention Module

The bi-directional attention module primarily identifies artifacts in the local facial components area. However, it ignores some additional global details, specifically those related to textures or edges, which are consistently found in high-frequency information, as displayed in Figure 3. To address this, we design a complementary attention module in the frequency domain (Figure 4).

This module first splits the input feature X into n parts $[X^0, X^1, \dots, X^{n-1}]$

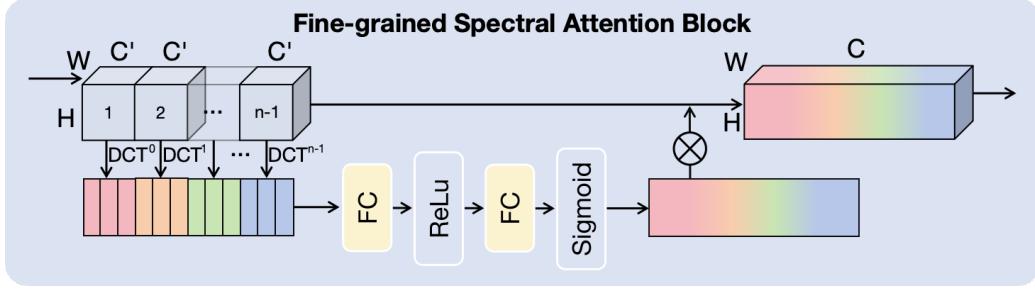


Figure 4: The structure of fine-grained spectral attention block.

among channel dimension, in which $X^i \in \mathbb{R}^{C' \times H \times W}$ ($C' = \frac{C}{n}, i \in [0, 1, \dots, n-1]$). Thereafter, we utilize the corresponding DCT [43, 44] to convert each input X^i into the frequency domain for the reason that DCT exhibits strong energy compaction properties [44]:

$$B_{h,w}^{u_i,v_i} = \cos\left(\frac{\pi h}{H}(u_i + \frac{1}{2})\right) \cos\left(\frac{\pi w}{W}(v_i + \frac{1}{2})\right)$$

$$\nu^i = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X_{:,h,w}^i B_{h,w}^{u_i,v_i}, \quad (9)$$

where $[u_i, v_i]$ are the frequency component indices corresponding to X^i , and $\nu^i \in \mathbb{R}^{C'}$ is the C' -th dimension vector after the compression. By doing so, the DCT is converged on a smaller scale, resulting in more reservation of high-frequency information. The whole compression spectral band κ can be obtained after concatenation operation $\epsilon[\cdot, \dots, \cdot]$:

$$\kappa = \epsilon([\nu^0, \nu^1, \dots, \nu^{n-1}]). \quad (10)$$

The weight κ' can be obtained with κ after different fully connected functions and activation functions, and the final output X_{sp} is:

$$X_{sp} = X \times \kappa'. \quad (11)$$

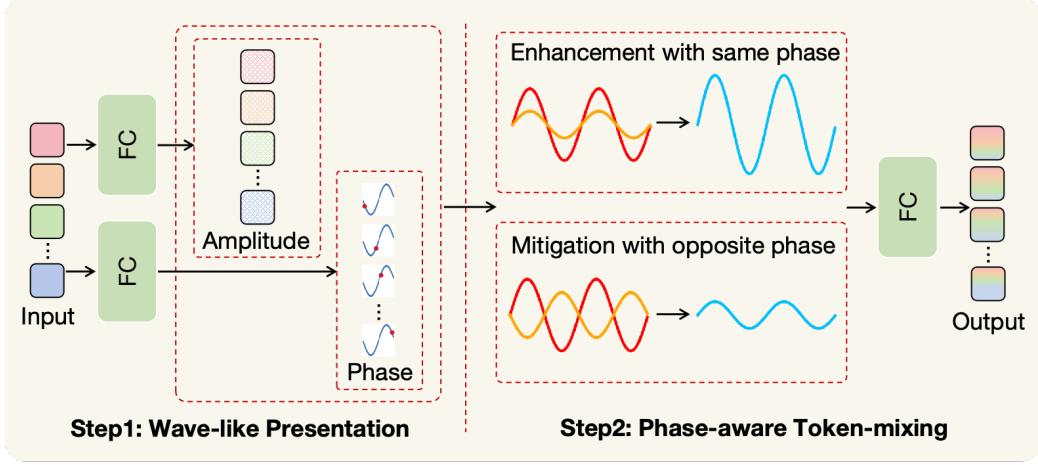


Figure 5: Two-step processing in feature superposition.

3.4. Feature Superposition

Although spatial features and frequency features can complement each other (e.g. the former contains the distribution of the local organs, the latter contains the global texture and edge information), the network still cannot effectively capture the feature differences between real and forgery. Therefore, we present the superposition strategy, which can enhance the fused features according to their positional information, thereby enabling the model to distinguish between various feature classes more effectively. Specifically, this strategy is divided into two stages: the implementation of a wave-like representation followed by phase-aware token-fusion (Figure 5).

3.4.1. Wave-like Representation.

After the merging feature $X' \in \mathbb{R}^{C \times H \times W}$ from $X_{bi} \in \mathbb{R}^{C \times H \times W}$ sum $X_{sp} \in \mathbb{R}^{C \times H \times W}$, we split X' into m tokens $[x'_0, x'_1, \dots, x'_{m-1}]$ among height and width dimension, and we generate the amplitude and phase of each token

$x_i' \in \mathbb{R}^{C \times \frac{H \times W}{m}}$. In our wave-like term, we assign each token ‘‘phase’’ and ‘‘amplitude’’ attributes, which form the foundation for subsequent feature processing. This enables us to dynamically integrate tokens based on their semantic information, amplifying the distances between tokens with different attributes and clustering tokens with similar attributes to achieve improved classification outcomes.

Amplitude: The amplitude refers to the local information contained in the token and the assigned global facial information. In our work, we generate amplitude information through a plain channel-FC operation ω_1 with the learnable parameters W^c , while the amplitude is a real-value feature, we employ an absolute value operation $|\cdot|$ to obtain the amplitude $|z_j|$ ($j \in [0, 1, \dots, m - 1]$):

$$|z_j| = |\omega_1(x_j', W^c)|. \quad (12)$$

Phase: The phase θ_j is the current location of the token within a wave period. In our work, we utilize channel-FC to extract the positional information of this token in the face, thereby concretizing the phase. Specifically, we adopt another simple channel-FC ω_2 with the learnable parameters W^q as the phase estimation module:

$$\theta_j = \omega_2(x_j', W^q). \quad (13)$$

Wave-like Representation: Each token is reinterpreted as a wave \tilde{z}_j , encapsulating both amplitude and phase information. For the purpose of embedding the wave-like token within a structure akin to that of the subsequent fully connected layers, the token undergoes expansion via the Euler formula

and is expressed through its real and imaginary components:

$$\tilde{z}_j = |z_j| \odot \cos\theta_j + a|z_j| \odot \sin\theta_j, \quad (14)$$

where a is the imaginary unit satisfying $a^2 = -1$, $|z_j| \odot \cos\theta_j$ is the real part, $a|z_j| \odot \sin\theta_j$ is the imaginary part.

3.4.2. Phase-aware Token-fusion

Tokens can be categorized into two groups: authentic and artifact. Current manipulated images are based on full-face region or local-face organs, without discrete generation of artifacts, so each category token has similar positional information formed as the phase.

When aggregating different tokens, supposing \tilde{z}_r is the resultant wave of \tilde{z}_k and \tilde{z}_j , its amplitude $|z_r|$ can be calculated as follows:

$$|z_r| = \sqrt{|z_k|^2 + |z_j|^2 + 2|z_k| \odot |z_j| \odot \cos(\theta_j - \theta_i)}. \quad (15)$$

The phase difference $(\theta_j - \theta_i)$ between two tokens has a significant effect on the amplitude of the aggregated result z_r . When two tokens have the same phase ($\theta_j = \theta_k + m\pi, m \in [0, \pm 2, \pm 4, \dots]$), they will be enhanced by each other, i.e., $|z_r| = |z_k| + |z_j|$. For the opposite phase ($\theta_j = \theta_k + m\pi, m \in [\pm 1, \pm 3, \dots]$), they will be weakened by each other, i.e., $|z_r| = ||z_k| - |z_j||$.

The tokens interact with each other with their own amplitude and phase, and iterated token \tilde{z}_r are then updated with the token-FC operation τ :

$$\begin{aligned} \tilde{o}_j &= \tau(\tilde{Z}, W^t)_j \\ &= \sum_r W_{jr}^t \odot \tilde{z}_r, \end{aligned} \quad (16)$$

where \tilde{o}_j is the updated tokens, $\tilde{Z} = [\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_n]$ denotes all the wave-like tokens in a layer, W^t is the token-fusion weight. Following common quantum

measurements [45, 46], we obtain the real-valued output o_j by reweighting and summing the real and imaginary parts of \tilde{o}_j .

$$o_j = \sum_r (W_{jr}^t z_r \odot \cos\theta_r + W_{jr}^i z_r \odot \sin\theta_r). \quad (17)$$

In the above equation, W_{jr}^t, W_{jr}^i are learnable weights, the phase θ_r is dynamically adjusted according to the semantic content of the input data. To increase the representation capacity, the final output P is obtained by transforming o_j with another Channel-FC operation ω_3 with weight W^p :

$$P = \omega_3(o_j, W^p). \quad (18)$$

4. Experimental Results

In this section, we first introduce the implementation details and present our experimental result on the FaceForensics++ [47], Celeb-DF [48], Celeb-DF-v2 [48], the Deepfake Detection Challenge [49], and DeeperForensics-1.0 [50]. Following this, we compare our proposed D²Fusion with other state-of-the-art methods, and an in-depth analysis is provided to understand our framework better.

4.1. Datasets and Experimental Settings

To evaluate the generalization ability of the proposed D²Fusion, we carry out the experiments on the following five public benchmark datasets:

FaceForensics++(FF++): The FF++ dataset is widely recognized as the most frequently used benchmark for detecting facial Deepfake videos. This dataset offers three compression rate options: original quality (RAW), high

quality with light compression (HQ), and low quality with heavy compression (LQ). FF++ is composed of four sub-datasets, each corresponding to a specific generation method. For generating FF++ videos, two of these methods are based on computer graphics approaches: Face2Face (F2F) and FaceSwap (FS), while the other two rely on deep-learning approaches: DeepFakes (DF) utilizing auto-encoder and NeuralTextures (NT) using GANs. Within the FF++ dataset, there are 1000 real videos and each sub-dataset generates an additional 1000 fake videos, resulting in a total of 5000 videos.

Celeb-DF (CD1): The CD1 dataset comprises of both original and synthesized videos that exhibit visual quality similar to the videos commonly encountered online. The Celeb-DF dataset encompasses 408 original videos sourced from YouTube, featuring subjects of diverse ages, ethnic backgrounds, and genders. Additionally, within the Celeb-DF dataset, there are 795 Deepfake videos generated through synthesis based on these original videos.

Celeb-DF-v2 (CD2): The CD2 dataset surpasses CD1 in terms of scale, containing 590 original videos along with 5639 corresponding Deepfake videos. Notably, the synthesized Deepfake videos within CD2 exhibit significant enhancements when compared to existing datasets. These improvements are particularly evident in areas such as stitching boundaries, color mismatches, inconsistent orientations, and other obvious visual artifacts [51].

Deepfake Detection Challenge (DFDC): The DFDC dataset represents a substantial dataset released for the Deepfake Detection Challenge compe-

Table 2: Publicly available datasets.

Dataset	FF++	CD1	CD2	DFDC	DFR
Real/Fake sample size	1000 real videos, 4000 fake videos	408 real videos, 795 fake videos	590 real videos, 5639 fake videos	19154 real videos, 100000 fake videos	50000 real videos, 10000 fake videos
Sample catagory	Face swapping Face reenactment	Face swapping	Face swapping	Face swapping	Face reenactment
Size before preprocessing	640×480, 1280×720, 1920×1080	Various	Various	320×240 - 3840×2160	1920×1080
Size after preprocessing				224×224	
Scenarios	YouTube	YouTube	YouTube	Actors	Actors

tition. This dataset comprises of 19,154 authentic videos sourced from 3,426 compensated actors, alongside 100,000 counterfeit videos generated through a range of Deepfake techniques. The authentic videos within the DFDC dataset closely mirror real-life scenarios, while the areas with artifacts in its forgery videos exhibit greater precision compared to other datasets.

DeeperForensics-1.0 (DFR): The DFR dataset is a high-quality dataset specifically crafted for real-world face forgery detection. It encompasses 60,000 videos, with a staggering 17.6 million frames, in this dataset, there are 48,475 source videos and 11,000 manipulated videos. The dataset places a strong emphasis on realism, scale, and diversity, capturing real-world variations such as sub-optimal illumination, extensive occlusion of the target faces, and extreme head poses. Concretely, in DFR, the authors proposed a many-to-many end-to-end face-swapping technique known as the Deepfake Variational Auto-Encoder [22] (DF-VAE) to generate Deepfake videos.

Implementation Details: We train D²Fusion on a single NVIDIA RTX3090

GPU and use PyTorch to build it. We adopt ResNet34 [4] as our backbone, which is pre-trained on the ImageNet dataset [52]. For each video, we evenly select 32 frames for testing and training. The Dlib is employed for face detection and extraction in the frames, and the cropped faces are resized to 224×224 . Multi-scale facial swap sliding window scale is randomly selected from $[40, 80]$, $[80, 120]$, $[120, 160]$, $[224, 224]$.

During the training phase, we set the batch size to 64 and record results at epoch 100, 200, 300, 400, 500 to find the optimal checkpoint. We use Adam [53] as our optimizer and the learning rate is set to 3.6×10^{-4} at initialization, and decrease to 2×10^{-4} at epoch 20, 1×10^{-4} at epoch 40, 5×10^{-5} at epoch 60, and 1×10^{-5} at epoch 80 for fine-tuning. It is worth noting that in our experiments, all evaluation measures are computed at the frame level.

Evaluation Metrics: In our different experiments, we utilize a total of five types five evaluation metric: the area under the curve (AUC), accuracy (ACC), precision(P), recall(R), and F1 score($F1$):

$$\begin{aligned} AUC &= p(p(\frac{TP}{TP+FN}) > p(\frac{FP}{TN+FP})), \\ ACC &= \frac{TP+TN}{TP+TN+FP+FN}, \\ P &= \frac{TP}{TP+FP}, \\ R &= \frac{TP}{TP+FN}, \\ F1 &= 2\frac{P \times R}{P + R}, \end{aligned} \tag{19}$$

$p(\cdot)$ denotes probability, TP , TN , FP and FN denote the counts of true

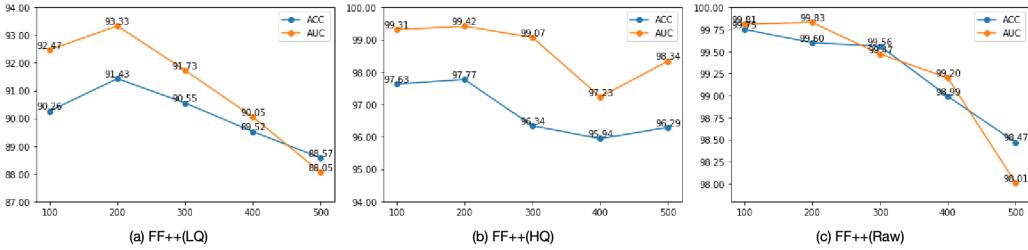


Figure 6: Incremental comparison experiments using three compression rate versions of FF++ at different epochs.

positive, true negative, false positive, and false negative samples, respectively.

4.2. Comparison with State-of-the-art Methods

We use various datasets to compare our method with other face forgery detection techniques to show its effectiveness and generalization performance. To make the comparison fair and complete, we reproduce some corresponding experiments using state-of-the-art (SOTA) methods under the same experiment setting. Note that in all results tables, we underline the second-best results while the best results of all listed methods are in bold.

4.2.1. Intra-dataset Evaluation on FF++

We initially assess D²Fusion at various epochs to determine the optimal checkpoint. In Figure 6, we conduct an incremental comparison and find that at epoch 200, the model achieves peak performance within the dataset across different compression rates. As the number of epochs increases, both ACC and AUC metrics show a decline.

We evaluate our method in comparison to previous detection methods using three versions of FF++. The comparative results are presented in Table 3. Overall, irrespective of the compression ratio applied to the video

Table 3: Intra-dataset evaluation with other methods on FF++ with different compression rate in terms of ACC (%) and AUC (%).

Method	Year	LQ		HQ		RAW	
		ACC	AUC	ACC	AUC	ACC	AUC
Face X-ray [6]	2020	-	61.60	-	87.35	-	-
GFF [33]	2021	86.89	88.27	96.87	98.95	-	-
MAT(EN-B4) [30]	2021	88.69	90.40	<u>97.60</u>	99.29	-	-
TAN-GFD [37]	2023	87.42	89.26	97.17	99.21	-	-
CADMM [32]	2023	89.04	92.32	97.59	99.24	<u>99.51</u>	<u>99.78</u>
FADE(ResNet) [31]	2023	<u>90.56</u>	<u>93.37</u>	97.55	99.31	-	-
IDM [36]	2024	-	93.57	-	99.74	-	-
Ours	2024	91.43	93.33	97.77	99.42	99.60	99.83

dataset, our proposed method consistently outperforms earlier visual artifact-based backbone networks such as Face X-ray, which can only detect the most basic face forgery operations.

In three versions of FF++, compared to the methods that use single domain information, in the case of raw quality images, Our method demonstrates a significant performance improvement, achieving a 0.09 increase in ACC and a 0.05 enhancement in AUC compared to CADMM [32]. Similarly, MAT [30] exhibits this deficiency with AUC decreased by 2.93. The superior performance of our approach contributes to its incorporation of the fine-grained spectral attention module, which effectively processes frequency domain information. Similarly, in the LQ version, D²Fusion surpasses FADE [31] by up to 0.22 in ACC and AUC metrics. As compared to FADE, which

Table 4: Cross-manipulation evaluation with other methods on FF++(HQ) in terms of AUC (%). In the *Bias* column, we show the AUC change when compared to CADDM.

Train	Test	EN-B4 [5] (2019)	MAT [30] (2021)	GFF [33] (2021)	DCL [54] (2022)	CADDM [32] (2023)	Ours	Bias
(Computer Graphics)	F2F	99.20	99.13	99.10	99.21	<u>99.67</u>	99.86	+0.19
	FS	58.14	60.14	61.30	59.58	64.07	<u>62.47</u>	-1.60
	DF	84.52	86.15	89.23	91.91	88.19	<u>89.50</u>	+1.31
	NT	63.71	64.59	64.77	66.67	<u>72.74</u>	75.23	+2.49
(Computer Graphics)	F2F	67.69	66.39	68.72	<u>69.95</u>	73.28	69.76	-3.52
	FS	99.89	99.67	99.85	99.90	<u>99.91</u>	99.92	+0.01
	DF	69.25	64.13	70.21	74.80	<u>75.66</u>	77.50	+1.84
	NT	48.61	50.10	49.91	52.60	59.94	<u>58.45</u>	-1.49
(Auto-encoder)	F2F	76.32	75.23	76.89	77.13	<u>77.85</u>	77.88	+0.03
	FS	46.24	40.51	47.21	61.01	64.93	<u>62.25</u>	-2.68
	DF	<u>99.97</u>	99.92	99.87	99.98	99.94	99.98	+0.04
	NT	72.72	71.08	72.88	<u>75.01</u>	71.86	75.73	+3.87
(GAN)	F2F	48.86	48.22	49.81	52.13	<u>67.03</u>	71.08	+4.05
	FS	73.05	75.33	74.31	<u>79.31</u>	74.60	80.75	+6.15
	DF	85.99	87.23	88.49	91.23	<u>92.40</u>	94.44	+2.04
	NT	98.25	98.66	98.77	98.98	<u>99.09</u>	99.43	+0.34

relies solely on a multi-dependency graph in the spatial domain, our approach leverages the fine-grained spectral attention module, enabling more effective feature extraction and performance gains. However, in the HQ version, D²Fusion exhibits slightly lower performance than IDM [36], with AUC decreased by 0.28. In the IDM model, it focuses on frequency domain information to decompose video frames into illumination and reflection components. This focus making it restrictive in very specific scenarios and thus resulting in suboptimal performance across different datasets. In Table 6, when training on FF++(HQ) and testing on CD1, the AUC value of the

IDM model decrease 11.6% compared to our method. Besides, the discrepancy in D²Fusion may be due to the loss of artifactual information in highly compressed videos, which also provides us with a future direction that the frequency domain might better preserve forged evidence in low-quality videos.

4.2.2. Cross-manipulation Evaluation on FF++(HQ)

In order to validate the generalization capability of our detection method, and demonstrate that our detection network is not restricted to detecting specific forgery generation techniques, we also evaluate the cross-manipulation performance of the proposed methods on FF++(HQ). We trained the involved models on one manipulation method and tested them on all methods with FF++(HQ). The result is shown in Table 4, obviously, the EfficientNet-B4 [30], MAT [30] and GFF [33] work well in response to the known forgery operation. However, their ability to distinguish between authentic and fake images is limited when faced with unfamiliar forgery patterns. It is should also be noted that, compared to the CADDM [32], our method demonstrated considerable improvements in both intra- and cross-manipulations with NT, with an increase of AUC 4.05%, 6.15%, 2.04%, and 0.34% on the F2F, FS, DF, and NT, respectively. This task is particularly challenging due to the NT dataset uses GANs to generate fake videos that manipulate the mouth expressions, resulting in more realistic and localized artifact regions. Notably, our method achieved a remarkable 6.15% improvement over the CADDM on the FS. It also outperformed the second-best method, DCL [54], by 1.44%, indicating the capability of D²Fusion to capture localized fake areas while addressing lower-level forgery manipulations. We also note that our proposed method exhibits slight performance regressions in certain scenarios, with a

Table 5: Multi-source manipulation evaluation with other methods on FF++(HQ) in terms of P , R , $F1$, ACC, AUC (%).

Method	Year	DF(Auto-encoder)					F2F(Computer Graphics)				
		P	R	$F1$	ACC	AUC	P	R	$F1$	ACC	AUC
MAT [30]	2021	81.47	78.95	80.19	80.54	85.94	72.63	78.54	75.47	75.02	78.52
CADDM [32]	2023	<u>86.17</u>	<u>89.97</u>	<u>88.02</u>	<u>87.79</u>	<u>91.13</u>	84.13	72.82	77.40	79.58	81.24
FreqNet [35]	2024	84.38	86.66	85.50	85.33	89.75	87.81	<u>80.94</u>	<u>84.24</u>	<u>84.89</u>	<u>86.32</u>
Ours	2024	87.09	94.08	90.45	90.11	95.34	<u>87.25</u>	82.35	84.73	85.70	87.23
Method	Year	FS(Computer Graphics)					NT(GAN)				
		P	R	$F1$	ACC	AUC	P	R	$F1$	ACC	AUC
MAT [30]	2021	74.70	62.19	67.87	70.63	72.56	<u>70.72</u>	65.40	67.96	64.21	68.20
CADDM [32]	2023	<u>76.61</u>	68.00	72.05	73.65	76.42	70.45	<u>70.31</u>	<u>70.38</u>	<u>70.45</u>	<u>74.50</u>
FreqNet [35]	2024	74.19	78.43	76.30	<u>75.64</u>	<u>80.23</u>	66.90	67.90	67.40	67.22	69.94
Ours	2024	79.64	<u>71.01</u>	<u>75.08</u>	76.47	81.99	78.92	71.72	75.15	76.34	80.26

3.52% decrease against CADDM when trained with FS and tested in F2F, and a 2.68% decrease when trained with DF and tested on FS. This could be attributed to an overrepresentation of artifact data type in the training set, which caused the model to overfit to a specific manipulation form. Overall, the results in Table 4 confirm that D²Fusion is effective in capturing artifact features across various forgery techniques by considering information from different domains.

4.2.3. Multi-source Manipulation Evaluation

Furthermore, we conduct a multi-source manipulation evaluation as an additional experiment to complement the cross-manipulation evaluation, given its inclusion of various manipulation operations. The performance of the proposed detection model under multi-source manipulation conditions is as-

Table 6: Cross-dataset evaluation on CD1, CD2, DFDC and DFR by training FF++(HQ) compared with other methods in terms of AUC (%).

Method	Year	CD1	CD2	DFDC	DFR
Face X-ray [6]	2020	80.58	74.20	70.00	-
MAT [30]	2021	-	67.44	67.34	-
GFF [33]	2021	-	76.65	71.58	-
TAN-GFD [37]	2023	-	72.33	73.46	-
CADDM [32]	2023	89.57	77.04	71.49	<u>86.92</u>
IDM [36]	2024	76.54	-	-	-
E-TAD [34]	2024	70.00	58.50	-	-
LSDA [38]	2024	86.70	<u>83.00</u>	<u>73.60</u>	-
Ours	2024	<u>88.14</u>	83.29	74.93	90.40

sessed on FF++(HQ). Specifically, the detection models are trained on three sub-datasets and tested on the remaining sub-dataset. As demonstrated in Table 5, the proposed method significantly surpasses MAT [30] in all the sub-datasets. Especially on the F2F dataset, the P value of our method exceeds MAT by 14.62, since MAT only focuses on spatial information. Similar issues arise with CADDM [32]; although it demonstrates relatively good generalization performance by extracting spatial features, its insufficient extraction of frequency domain information results in a 9.47% reduction in the F1 score on the F2F dataset. FreqNet [35] performs well in F2F and FS datasets, which use computer graphics to generate Deepfake videos, but FreqNet underperforms against more advanced adversarial attacks such as those from Auto-encoders or GANs due to its bias towards specific high-level frequency information. Particularly, on the NT dataset, our method surpasses FreqNet by 12.02, 3.82, and 7.75 in P , R , $F1$ respectively.

4.2.4. Cross-dataset Evaluation

Although the aforementioned FF++ sub-datasets utilize four distinct forgery algorithms, the fake videos originate from identical source videos. To evaluate the performance of more different data distributions, we use the FF++(HQ) data set as the training data set and evaluate other data sets. From Table 6, it becomes evident that, when comparing our framework against various mainstream detection methods, our proposed D²Fusion significantly outperforms earlier basic reference methods, such as Face X-ray [6]. Compared to methods utilizing a single domain, such as MAT [30] and GFF [33], our approach demonstrates significant advantages. On the CD2 dataset, our method exceeds MAT by 23.50%, and on the DFDC dataset, it surpasses GFF by 3.92%. Notably, our method registers performance improvements of 8.11% and 2.00% over the advanced models TAN-GFD [37] in CD2 and DFDC respectively. The AUC performance of our network exhibits a 4.00% improvement over our baseline CADMM in DFR. Compared to most recent works [36, 34], our method outperforms E-TAD, IDM that only uses the frequency domain. For example, in CD1, our method exceeds E-TAD by 40.2%; in DFDC, our method exceeds LSDA [38] by 1.33. This improvement arises because our method delves deeper into feature exploration than LSDA. These results affirm that D²Fusion adeptly captures artifacts with greater precision, even amidst previously unseen backgrounds or identities.

In summary, our proposed D²Fusion outperforms previous SOTA techniques, showcasing its robustness and effectiveness in detecting forged content across a wide range of manipulation techniques and real-life scenarios.

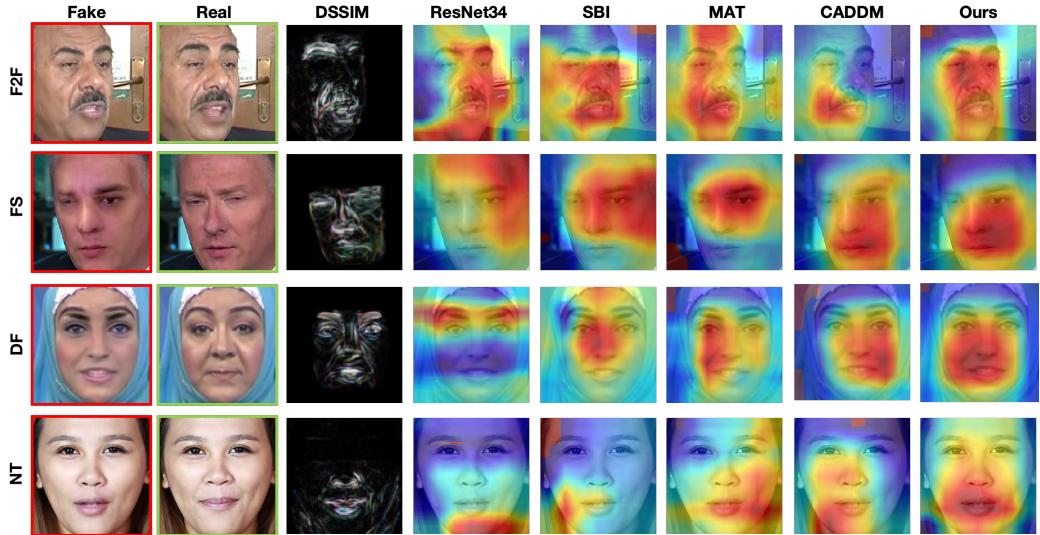


Figure 7: Qualitative analysis on FF++(HQ). We show our detection results with multi-source manipulation evaluation. DSSIM images indicate the forgery area, and we can see that D²Fusion can better capture the manipulated area compared to other main-stream detection methods.

4.3. Qualitative Results

The qualitative results are displayed in Figure 7. We use the Equation (1) to draw the DSSIM pictures to help intuitively identify tampered areas. Additionally, we employ Grad-CAM [55] for generating heat maps, which visualize the focus regions of the models by fusing weights into the gradient information. In addition to vanilla ResNet34, we chose MAT [30], SBI [7], and CADDM [32] as our comparison methods. To test whether these models are effective to capture artifact features between different forgery methods, we perform multi-source manipulation experiments results.

From the heat maps of Resnet34, it becomes clear that this vanilla CNN are capable of detecting obvious blending boundary, which is a type of low-

level feature easily perceptible even to the human eye. The detection performance is satisfactory for images with these evident clues in DF and FS, yet it proves ineffective for F2F and NT. Especially in NT, a facial reenactment dataset that alters expressions while preserving facial contours, ResNet34 still predominantly focuses on the boundary area.

SBI introduces self-blending images (the basis for multi-scale facial swap) for data augmentation, employing EfficientNet-B4 [5] as the detection network. As depicted in Figure 7, although SBI approximately identifies forgery areas, its accuracy is lacking. In F2F, FS, and NT, SBI inaccurately classifies some genuine areas as manipulated. In contrast to regular ResNet34, SBI enriches data diversity through self-blending, expanding detection beyond merely the edge regions. However, the inherent detection capabilities of the model remain limited.

In the case of MAT, the detection model employs multiple spatio attention heads. As illustrated in Figure 7, the model is capable of focusing on various local parts. However, it fails to identify the general forgery area in these four example images accurately. Specifically, in NT, the area with the most significant mismatch between detected artifacts and actual forgeries is observed. This limitation stems from the exclusive reliance of MAT on spatial information, without incorporating data from other domains.

Compared to SBI and MAT, our baseline CADDMM demonstrates precise alignment in detection, accurately pinpointing areas due to its utilization of multi-scale detection for local feature extraction. However, in F2F and DF, it does not capture all manipulated regions, suggesting that there is room for improvement in global feature learning. Our detection outcomes

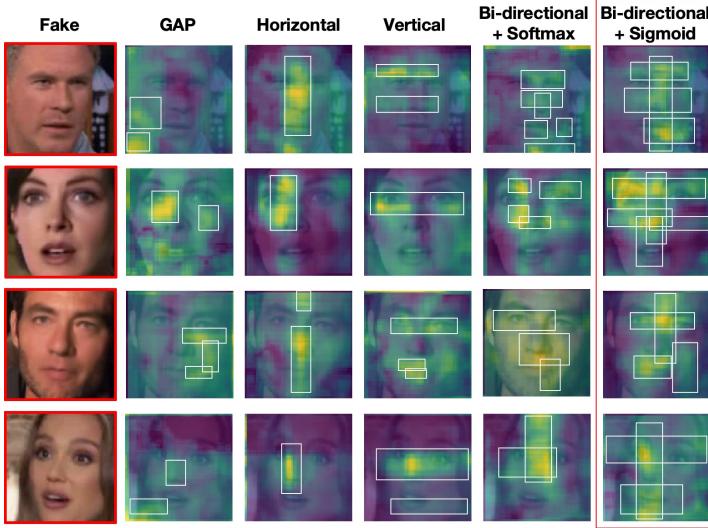


Figure 8: From left to right are the fake images and the feature maps after GAP, horizontal average pooling, vertical average pooling, bi-directional average pooling with Softmax function and bi-directional average pooling with Sigmoid function. It should be noted that the forged features are marked by the white bounding boxes.

reveal that D²Fusion surpasses the baseline in effectiveness. In F2F, FS and DF, the localization closely matches the forged area, accurately encompassing a more extensive portion of the forgery compared to the baseline. In NT, the detection more accurately focuses on the forged mouth area. These results conclusively show that our model can efficiently identify artifacts, independent of the forgery techniques or whether the manipulation is global or localized.

4.4. Ablation Studies

4.4.1. Evaluation on Components Intrinsic

We particularly visualize the intrinsic effects of individual components on model performance. Figure 8 demonstrates that bi-directional average pool-

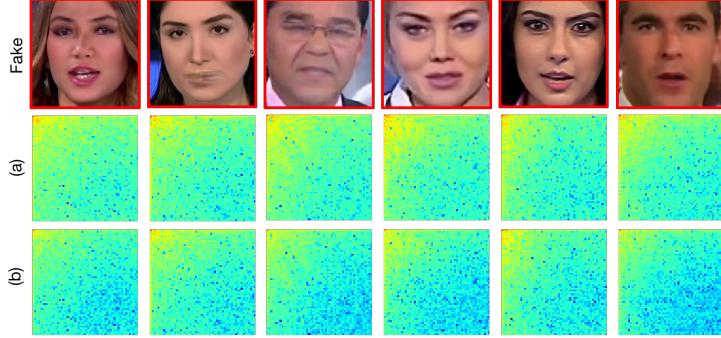


Figure 9: (a) displays the log-scaled spectrum images after using a single DCT over the entire channel, (b) displays the log-scaled spectrum images after using multi-spectral partitioning and the respective corresponding DCTs.

ing can accurately localize manipulated regions compared to global average pooling (GAP). It can also effectively preserve the comprehensive spatial distribution of artifact clues from both directions compared to uni-directional average pooling. The visualization further reveals that using the Sigmoid function enables more concentrated and complete localization of artifact features. In contrast, while the Softmax function can also process features via bi-directional average pooling operations, its mapping results in a more dispersed localization of artifact features.

Figure 9 (a) illustrates that using a single DCT over the entire channel, the spectrum contains lower frequency components, while the spectrum band calculated from Equation (9) and Equation (10) can retain stronger high-frequency components as shown in Figure 9 (b).

Figure 10 shows the error maps before and after feature superposition. It is obvious that with the intervention of feature superposition, the difference between the authentic features and the artifact features within the image

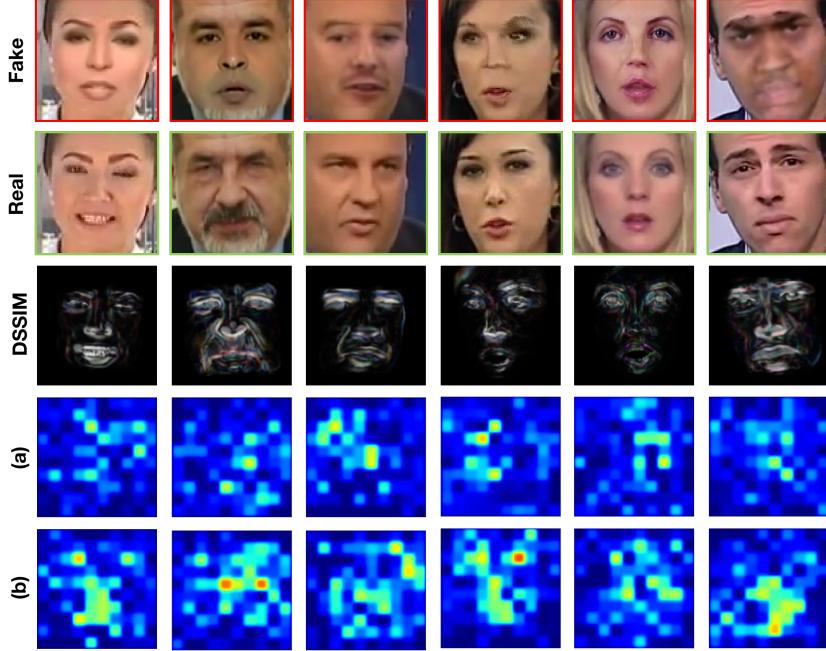


Figure 10: (a) displays the error maps of $|f_f - f_r|$, in which f_f is the fake image feature, f_r is the real image feature. (b) displays the error maps of $|f'_f - f_r|$, in which f'_f is the same fake image feature after feature superposition processing.

becomes more significant, which is manifested as a brighter region.

4.4.2. Evaluation on Individual Component

Table 7 and Table 8 reveal that the integration of a single domain module into vanilla ResNet34 proves to be effective. Specifically, as shown in Table 7, the addition of the fine-grained spectral module results in a 1.70 increase in R for experiments within the dataset, indicating that the model is able to identify artifact samples more comprehensively. In cross-dataset experiments, DFR experiences the most significant improvement, with a 6.06% rise in AUC. The incorporation of the fine-grained spectral attention module leads to AUC enhancements of 1.63% in FF++, 2.10% in CD1, 2.40% in

Table 7: Ablation study on the effect of different components of our model using five metrics (%) in intra-dataset evaluation with FF++(HQ).

Bi-directional	Fine-grained	Spectral	Feature Superposition	FF++				
				P	R	F1	ACC	AUC
-	-	-	-	97.23	95.09	96.14	96.20	97.25
✓	-	-	-	97.43	94.98	96.19	96.22	98.79
-	✓	-	-	96.02	96.79	96.40	96.41	98.45
-	-	✓	-	96.76	95.79	96.27	96.28	97.99
-	✓	✓	✓	96.52	<u>97.29</u>	96.90	96.88	98.99
✓	-	✓	✓	96.17	98.40	97.27	97.24	99.13
✓	✓	-	-	<u>97.98</u>	<u>97.29</u>	<u>97.63</u>	<u>97.65</u>	<u>99.20</u>
✓	✓	✓	✓	98.08	97.39	97.73	97.77	99.42

CD2, 3.07% in DFDC, and 7.78% in DFR. Implementing the feature superposition strategy yields R , $F1$, ACC, and AUC improvement of 0.70, 0.13, 0.08, 0.74 in intra-dataset scenarios, respectively, and a maximum increase AUC of 10.61% in cross-dataset scenarios with DFR.

We utilize visualization tools to highlight the critical role of our designed attention modules. As depicted in Figure 11 (a) for the intra-dataset scenario, it becomes evident that models lacking the bi-directional module can identify tampered areas, yet with insufficient precision, particularly in FS. This situation suggests that spatial information could be more effectively processed. Conversely, models missing the fine-grained spectral attention module might accurately locate tampered zones but fail to encompass all manipulated areas as comprehensively as desired. Such findings indicate the underutilization of frequency domain information, signalling opportunities for enhancement. In the heat maps of our model, it is observed that the at-

Table 8: Ablation study on the effect of different components of our model using AUC (%) metric in cross-dataset evaluation.

Bi-directional	Fine-grained Spectral	Feature Superposition	CD1	CD2	DFDC	DFR
-	-	-	83.43	79.27	67.34	69.50
✓	-	-	84.65	<u>82.73</u>	71.67	75.56
-	✓	-	85.53	81.67	70.41	77.28
-	-	✓	84.37	82.72	69.20	80.11
-	✓	✓	<u>86.40</u>	82.42	72.88	82.19
✓	-	✓	86.17	81.09	72.37	<u>87.42</u>
✓	✓	-	85.81	82.54	<u>73.46</u>	85.80
✓	✓	✓	88.14	83.29	73.93	90.40

tention mechanisms enable the model to localize the manipulation area with greater precision, regardless of the manipulation technique employed.

In Figure 11 (b), we present visualizations for cross-dataset scenarios. The outcomes from models lacking bi-directional attention reveal instances of positioning overflow, notably in CD1 and CD2. Similarly, models devoid of fine-grained spectral attention tend to identify smaller forgery areas across datasets. In stark contrast, our comprehensive network showcases remarkable robustness, precisely locating and identifying manipulated regions, regardless of the complexity of real-world scenarios or the diversity of ID representations. This demonstrates that adding the attention components enhances the ability of the model to capture the abnormal regions and significantly improves its adaptability in dealing with variable and unknown pattern modifications. This phenomenon also accords with our motivation.

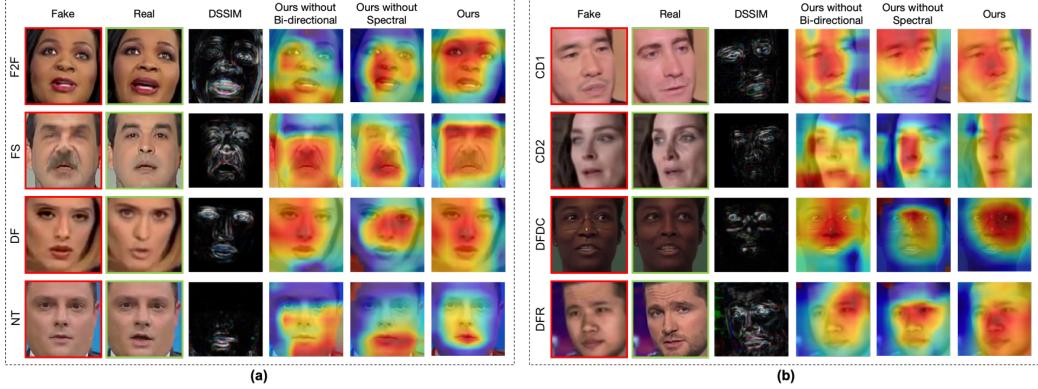


Figure 11: The Grad-Cam visualizations of samples in different datasets. Sub-figure (a) exhibits intra-dataset experiments, and sub-figure (b) exhibits cross-dataset experiments. The heat maps prove that with the introduction of our attention modules for different forgery methods and real-word information, D²Fusion can more precisely capture the forgery region.

Finally, we apply t-SNE [56] to visualize feature vector distribution in testing dataset from the last layers of the models. The visual results, as depicted in Figure 12, reveal that the feature distribution boundary appears relatively unambiguous in models lacking the feature superposition step, especially in CD1 and CD2. In contrast, the different categories of samples after the feature superposition process are further separated in the t-SNE embedding space, thus reflecting that introducing our feature superposition can help the model better increase the clarity of decision-making boundaries. This phenomenon is also reflected in Table 8, where within our complete framework, the introduction of feature superposition causes the AUC to rise by 0.09% in FF++(HQ), 2.33% in CD1, 0.75% in CD2, 0.47% in DFDC, 4.60% in DFR.

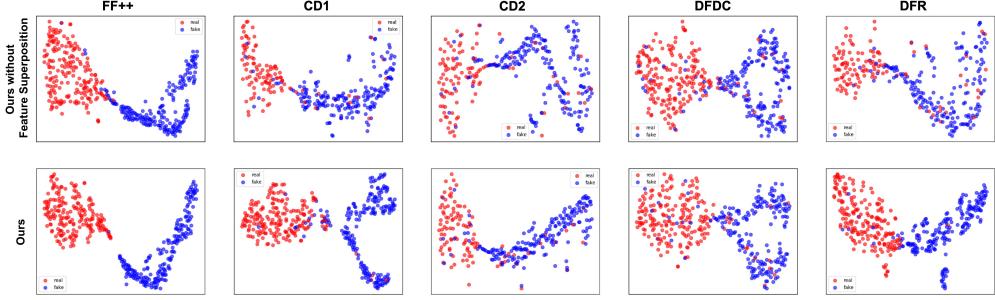


Figure 12: Visualization of t-SNE feature embedding in different scenarios, note that the red dots represent the real samples, while the blue dots represent the fake samples.

4.5. Evaluation of Network Load

To illustrate network load, Table 9 presents the total number of training parameters, which measure the size of models, and Floating Point Operations per Second (FLOPs) to assess the complexity of models. In detection within the FF++(HQ) dataset, D²Fusion outperforms all approaches in Table 9 in terms of accuracy with a relatively small network overload increase.

It is evident that baseline networks such as EfficientNetB4 [57], Xception [3], and ResNet34 [4] have fewer parameters and FLOPs because these networks serve solely as classification models without any architectural reconfiguration or additional modules tailored. This limitation results in poorer detection performance in deepfake detection tasks. In networks using EfficientNetB4 as the backbone, the MAT [30], and TAN-GFD [37] models see an increase in parameters by 5.96M and 13.39M, respectively. The former arises from introduced attention modules, while the latter incorporates the processing of multi-scale texture difference features. In networks with Xception as the backbone, GFF [33] significantly increases 5.89G in FLOPs due

Table 9: Network overload comparison in terms of Parameters(M) and FLOPs(G) with ACC in intra-dataset evaluation on FF++(HQ). In () represents the difference compared to the corresponding backbone.

Model	Parameters(M)	FLOPs(G)	ACC
EfficientNetB4 [57]	33.48	1.60	96.13
MAT(EfficientNetB4-based) [30]	39.44(<u>5.96↑</u>)	2.17(<u>0.57↑</u>)	97.60(<u>1.47↑</u>)
TAN-GFD(EfficientNetB4-based) [37]	46.87(13.39↑)	4.34(2.74↑)	97.17(1.04↑)
Xception [3]	39.70	4.61	95.49
GFF (Xception-based) [33]	53.25(13.55↑)	10.50(5.89↑)	96.87(1.38↑)
ResNet34 [4]	37.19	3.68	96.20
CADDM(ResNet34-based) [32]	40.60(3.41↑)	4.08(0.40↑)	<u>97.59</u> (1.39↑)
Ours(ResNet34-based)	45.66(8.47↑)	5.97(2.29↑)	97.77 (1.57↑)

to its dual-stream architecture designed to process both spatio-frequency features. However, although D²Fusion also addresses spatio-frequency features and has more backbone parameters than those in GFF, the overall parameter for our model is substantially less than that of GFF. Compared to CADDM, D²Fusion significantly enhances generalization performance with a relatively small network overload increase of 1.89G more FLOPs and exceeds CADDM by 0.18 ACC. For its backbone ResNet34, although D²Fusion incurs an additional 2.29G FLOPs, it achieves the most significant increase in accuracy, with an improvement of 1.57.

5. Limitation

As Table 3 demonstrates, our method exhibits a decrease in performance when dealing with high compression rate videos, in other words, low-quality

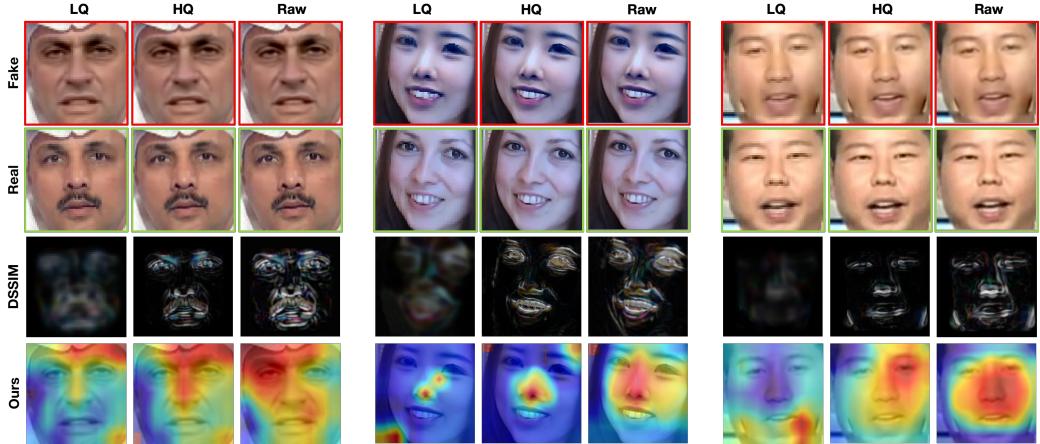


Figure 13: Failure cases. These samples all fail to be detected under low-quality video conditions, but they are successfully detected in high-quality and raw videos.

videos. The primary cause of this decline is the high compression rate, causing the loss of forgery clues within the video. Figure 13 clearly illustrates the impact of video compression rate on detection outcomes. In Figure 13, the human eyes may struggle to differentiate images at varying compression rates, but the DSSIM visualization demonstrates how an increased compression rate results in a decline in the loss of artifact information. Consequently, D²Fusion erroneously focuses on areas unrelated to the forged regions during low-quality video detection. As the compression rate declines and video quality improves, the increase in forgery details prompts the model to adjust its area of focus, thereby correcting the detection outcomes.

6. Future work

As discussed in Section 5, video quality can interfere with detection outcomes, and in real life, detectors often encounter batches of low-quality

videos, due to the varying levels of compression applied during the transmission through social media networks. Therefore, detectors need to enhance their ability to handle high-compression rate videos in the future. To this end, the following optimization directions are proposed:

- We will persistently refine and deepen our research on capturing local positional information of artifact clues from the spatial domain. This effort aims to achieve more precise localization and extend our research findings to practical applications.
- We plan to explore the impact of various frequency compression methods, such as wavelet transforms, on the preservation of forgery clues. Through this investigation, we aim to gain a deeper understanding of the effectiveness of different frequency compression technologies in maintaining original forgery features.
- We plan to incorporate recent datasets focused on highly compressed videos to explore low-quality video detection. This direction aims to adapt to the properties of low-quality videos, thereby enhancing the performance and reliability of detectors in practical media.

7. Conclusion

Existing works do not take into account the local and global properties of Deepfake videos in different domains. They also lack consideration for addressing the fundamental interaction issues in between different domains by utilizing the local and global complementarity in the spatial and frequency domain, respectively. To address the problem of insufficient feature processing and interaction in Deepfake detection, in this work, we introduce a novel

facial manipulation detection network, termed D²Fusion. Specifically, we first introduce a bi-directional attention module designed to average features horizontally and vertically. This module effectively maintains the spatial distribution of artifacts, thus improving the precision of artifact localization. To better extract sufficiently detailed artifact information, we utilize a fine-grained frequency attention module. By preserving the high-frequency components, this module significantly captures more of the details in the artifact features. In addition, our designed feature superposition strategy accepts two domain features to amplify the difference between authentic and artifact features, thus helping the detection model to be more generalized across different manipulation operations and real-world scenarios.

We evaluate our detection network in various experimental scenarios with significantly enhanced detection capabilities. In intra-dataset evaluation, D²Fusion exceeds CADDM [32] by 0.18 ACC within FF++(HQ). Furthermore, the model exhibits strong generalization capabilities in cross-dataset evaluations. On CD1, our method surpasses the emerging LSDA [38] by 1.44 in AUC, and this is critical for practical applications where the model must perform well on previously unseen techniques and real-life scenarios.

However, our work acknowledges limitations in the performance of D²Fusion model under high video compression. In the future, we propose further enhancements for handling such videos by improving local artifact detection and different frequency transforms [58, 59] with advanced low-quality video datasets. Our work aims to enhance deepfake detection techniques for practical implementation, especially targeting the identification of manipulated content in low-quality videos frequently transmitted through social media

platforms.

References

- [1] D. Yadav, S. Salmani, Deepfake: A survey on facial forgery technique using generative adversarial network, in: 2019 International conference on intelligent computing and control systems, IEEE, 2019, pp. 852–857.
- [2] 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.
- [3] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
- [4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [5] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.
- [6] 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, 2020, pp. 5001–5010.

- [7] K. Shiohara, T. Yamasaki, Detecting deepfakes with self-blended images, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18720–18729.
- [8] M. Yu, H. Li, J. Yang, X. Li, S. Li, J. Zhang, Fdml: Feature disentangling and multi-view learning for face forgery detection, Neurocomputing (2023) 127192.
- [9] R. Durall, M. Keuper, F.-J. Pfreundt, J. Keuper, Unmasking deepfakes with simple features, arXiv:1911.00686 (2019).
- [10] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, Leveraging frequency analysis for deep fake image recognition, in: International conference on machine learning, PMLR, 2020, pp. 3247–3258.
- [11] Y. Jeong, D. Kim, Y. Ro, J. Choi, Freqgan: robust deepfake detection using frequency-level perturbations, in: Proceedings of the AAAI conference on artificial intelligence, volume 36, 2022, pp. 1060–1068.
- [12] A. Agarwal, A. Agarwal, S. Sinha, M. Vatsa, R. Singh, Md-csdnetwork: Multi-domain cross stitched network for deepfake detection, in: 2021 16th IEEE international conference on automatic face and gesture recognition (FG 2021), IEEE, 2021, pp. 1–8.
- [13] Y. Wang, K. Yu, C. Chen, X. Hu, S. Peng, Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7278–7287.

- [14] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, N. Yu, F 2 trans: High-frequency fine-grained transformer for face forgery detection, *IEEE Transactions on Information Forensics and Security* 18 (2023) 1039–1051.
- [15] P. Yu, Z. Xia, J. Fei, Y. Lu, A survey on deepfake video detection, *Iet Biometrics* 10 (2021) 607–624.
- [16] Y. Lin, Q. Lin, F. Tang, S. Wang, Face replacement with large-pose differences, in: *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1249–1250.
- [17] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, G. Medioni, On face segmentation, face swapping, and face perception, in: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, 2018, pp. 98–105.
- [18] B. M. Smith, L. Zhang, Joint face alignment with non-parametric shape models, in: *Computer Vision–ECCV 2012: 12th European conference on computer vision, Florence, Italy, October 7–13, 2012, Proceedings, Part III* 12, Springer, 2012, pp. 43–56.
- [19] M. S. Rana, M. N. Nobi, B. Murali, A. H. Sung, Deepfake detection: A systematic literature review, *IEEE access* 10 (2022) 25494–25513.
- [20] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Faceshifter: Towards high fidelity and occlusion aware face swapping, *arXiv preprint arXiv:1912.13457* (2019).

- [21] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, et al., Deepfacelab: Integrated, flexible and extensible face-swapping framework, arXiv preprint arXiv:2005.05535 (2020).
- [22] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
- [23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2018, pp. 8789–8797.
- [24] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.
- [25] Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7184–7193.
- [26] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, arXiv preprint arXiv:1811.00656 (2018).
- [27] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, L. Verdoliva, Id-reveal: Identity-aware deepfake video detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 15108–15117.

- [28] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, B. Guo, Protecting celebrities from deepfake with identity consistency transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9468–9478.
- [29] K. K. Bhaumik, S. S. Woo, Exploiting inconsistencies in object representations for deepfake video detection, in: Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes, 2023, pp. 11–15.
- [30] 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, 2021, pp. 2185–2194.
- [31] L. Tan, Y. Wang, J. Wang, L. Yang, X. Chen, Y. Guo, Deepfake video detection via facial action dependencies estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 5276–5284.
- [32] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, Z. Ge, Implicit identity leakage: The stumbling block to improving deepfake detection generalization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 3994–4004.
- [33] 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, 2021, pp. 16317–16326.

- [34] J. Gao, M. Micheletto, G. Orrù, S. Concas, X. Feng, G. L. Marcialis, F. Roli, Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection, *Engineering Applications of Artificial Intelligence* 133 (2024) 108450.
- [35] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, Y. Wei, Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024, pp. 5052–5060.
- [36] C. Zhu, B. Zhang, Q. Yin, C. Yin, W. Lu, Deepfake detection via inter-frame inconsistency recomposition and enhancement, *Pattern Recognition* 147 (2024) 110077.
- [37] Y. Zhao, X. Jin, S. Gao, L. Wu, S. Yao, Q. Jiang, Tan-gfd: generalizing face forgery detection based on texture information and adaptive noise mining, *Applied Intelligence* (2023) 1–21.
- [38] Z. Yan, Y. Luo, S. Lyu, Q. Liu, B. Wu, Transcending forgery specificity with latent space augmentation for generalizable deepfake detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8984–8994.
- [39] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, Repvgg: Making vgg-style convnets great again, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13733–13742.

- [40] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE transactions on image processing* 13 (2004) 600–612.
- [41] A. Loza, L. Mihaylova, N. Canagarajah, D. Bull, Structural similarity-based object tracking in video sequences, in: 2006 9th International Conference on Information Fusion, IEEE, 2006, pp. 1–6.
- [42] Q. Hou, L. Zhang, M.-M. Cheng, J. Feng, Strip pooling: Rethinking spatial pooling for scene parsing, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4003–4012.
- [43] N. Ahmed, T. Natarajan, K. R. Rao, Discrete cosine transform, *IEEE transactions on Computers* 100 (1974) 90–93.
- [44] K. R. Rao, P. Yip, Discrete cosine transform: algorithms, advantages, applications, Academic press, 2014.
- [45] V. B. Braginsky, F. Y. Khalili, Quantum measurement, Cambridge University Press, 1995.
- [46] K. Jacobs, D. A. Steck, A straightforward introduction to continuous quantum measurement, *Contemporary Physics* 47 (2006) 279–303.
- [47] 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, 2019, pp. 1–11.

- [48] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3207–3216.
- [49] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, arXiv preprint arXiv:2006.07397 (2020).
- [50] L. Jiang, R. Li, W. Wu, C. Qian, C. C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2889–2898.
- [51] A. V. Nadimpalli, A. Rattani, On improving cross-dataset generalization of deepfake detectors, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 91–99.
- [52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE/CVF conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
- [53] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- [54] 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, volume 36, 2022, pp. 2316–2324.

- [55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
- [56] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).
- [57] B. Koonce, B. Koonce, Efficientnet, Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization (2021) 109–123.
- [58] H. Duan, Y. Long, S. Wang, H. Zhang, C. G. Willcocks, L. Shao, Dynamic unary convolution in transformers, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 12747–12759.
- [59] X. Miao, H. Duan, Y. Bai, T. Shah, J. Song, Y. Long, R. Ranjan, L. Shao, Laser: Efficient language-guided segmentation in neural radiance fields, arXiv preprint arXiv:2501.19084 (2025).