IDCNet: Image Decomposition and Cross-View Distillation for Generalizable Deepfake Detection

Zhiyuan Wang, Yanxiang Chen, *Member, IEEE*, Yuanzhi Yao, *Member, IEEE*,
Meng Han, *Senior Member, IEEE*, Wenpeng Xing, and Meng Li, *Senior Member, IEEE*

***Abstract*—Existing deepfake detectors predominantly process entire facial images as input, which limits their sensitivity to local forgery cues due to representation bias and information loss through CNN feature aggregation. To address these limitations, we propose IDCNet, a novel deepfake detection framework based on image decomposition and cross-view distillation. Our key insight is that decomposing images into complementary views enables specialized processing of global and local forgery cues, while cross-view distillation facilitates their mutual enhancement. Specifically, the framework employs a lightweight U-Net generator with a dual-objective mechanism to decompose input images into global content and local detail views, optimized through reconstruction and classification losses. A cross-view distillation strategy is then applied to enhance complementary feature learning between views. Furthermore, to integrate local artifact information into existing detection models without architectural modifications, we propose a feature alignment method. Extensive experiments across 14 forgery methods demonstrate the effectiveness of our approach, achieving up to 4.4% AUC improvement on the CDFV2 dataset compared to state-of-the-art methods. The source code is available at: https://github.com/wangzhiyuan120/idcnet**

***Index Terms*—Face forgery detection, deepfake detection, multi-view learning, representation disentanglement, mutual information.**

I. Introduction

In RECENT years, the rapid evolution of artificial intelligence (AI), particularly the advent of sophisticated generative models such as Generative Adversarial Networks (GANs) [1], [2], [3] and diffusion models [4], [5], has facilitated remarkable advancements in synthetic facial image manipulation. These technological innovations enable the generation of highly convincing deepfake content, permitting modifications to an individual's identity, physiological characteristics, and distinctive attributes. Deepfake technology, which uses artificial intelligence to create realistic but fabricated video and audio, has significant ethical consequences. Its misuse can lead to serious problems like identity theft, financial fraud, and breaches of biometric security systems [6], [7], [8], [9]. These issues can reduce public trust in the truthfulness of digital media. Therefore, it is a critical research goal to develop ways to counteract the harmful use of deepfakes.

Early approaches to deepfake detection relied on handcrafted features like eye blinking patterns and head pose inconsistencies. The field subsequently evolved toward deep learning-based methods that frame detection as a binary classification task, achieving impressive performance in controlled settings [10], [11]. However, these approaches often fail to generalize when confronted with novel manipulation techniques or varying image conditions [12], [13].

Recent efforts to improve generalization have explored various directions including data augmentation [14], [15], frequency domain analysis [16], [17], and identity preservation [13], [18]. However, existing approaches commonly process entire facial images as input. This prevalent paradigm inherently struggles to capture local manipulation artifacts, which are crucial for reliable deepfake detection. The limitations of this approach are twofold. First, conventional convolutional neural networks (CNNs) often exhibit a representational bias toward prominent global features (such as facial structure or identity) [13], which may cause them to overemphasize broad patterns while neglecting subtle, localized artifacts that are critical for accurate forgery detection. Second, the hierarchical feature aggregation mechanisms in CNNs—such as pooling and striding—can result in the irreversible loss of fine-grained spatial details and high-frequency information. These subtle features often carry key indicators of manipulation and are difficult to recover once discarded [19], [20].

To address these limitations, we propose IDCNet, a novel deepfake detection framework that explicitly extracts and integrates global and local forgery features through image decomposition and cross-view distillation. The main objective is to mitigate the global feature bias of CNNs and reduce the loss of fine-grained spatial details by leveraging complementary representations of the input image. As illustrated in Fig.1, unlike conventional methods that process the entire

Received 6 December 2024; revised 9 June 2025; accepted 22 July 2025. Date of publication 28 July 2025; date of current version 12 August 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62372149 and Grant U23A20303 and in part by the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China) under Grant BigKEOpen2025-04. The associate editor coordinating the review of this article and approving it for publication was Prof. Benedetta Tondi. *(Corresponding author: Yanxiang Chen.)*

Zhiyuan Wang is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230002, China (e-mail: zhiyuanwang@mail.hfut.edu.cn).

Yanxiang Chen, Yuanzhi Yao, and Meng Li are with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, the School of Computer Science and Information Engineering, and the Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei University of Technology, Hefei 230002, China (e-mail: chenyx@hfut.edu.cn).

Meng Han and Wenpeng Xing are with Binjiang Institute of Zhejiang University, Hangzhou 310027, China (e-mail: mhan@zju.edu.cn; xingwenpeng@zju-bj.com).

Digital Object Identifier 10.1109/TIFS.2025.3593353

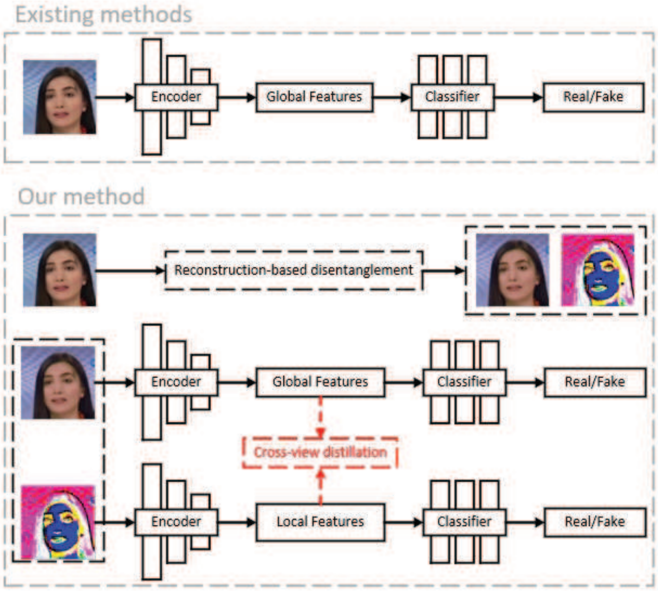


Fig. 1. Comparison between existing methods and our proposed approach. While existing methods analyze entire images, our framework decomposes facial inputs into complementary global content and local detail views. Cross-view distillation enables bidirectional knowledge transfer between these components for enhanced forgery detection.

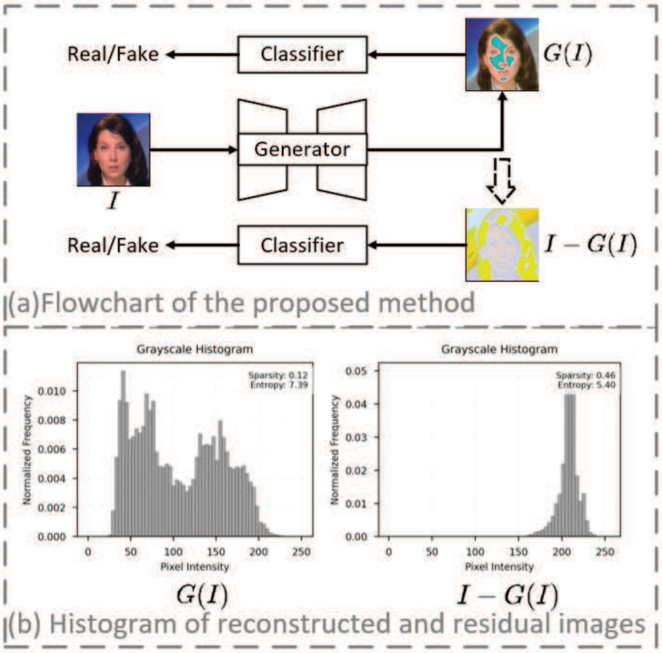


Fig. 2. (a) The workflow of our image decomposition process. (b) Visualization of the decomposition results. Our dual-objective U-Net separates facial images into global content and local detail components, guided by two classification networks. The decomposition achieves effective separation of structural information and high-frequency details.

image directly, IDCNet decomposes the input facial image into two complementary components: a global content image, which preserves the overall color and structural information, and a local detail image, which emphasizes edge features and high-frequency artifacts. This decomposition enables targeted learning of local forgery traces without interference from dominant global features. To further address the information loss inherent in hierarchical feature extraction, IDCNet introduces a cross-view distillation mechanism that facilitates bidirectional knowledge exchange between the global and local branches. This design ensures that fine-grained local details are preserved and enhanced, enabling more effective detection by jointly leveraging global and local manipulation cues.

Specifically, to achieve effective image-level decomposition, we employ a lightweight generator (U-Net [21] architecture) inspired by representation learning methods [22], [23], [24], [25]. The key insight is to precisely control the generator's reconstruction capability to separate the input facial image into two complementary components: a global content image preserving semantic structure and a local detail image capturing manipulation traces. As shown in Fig.2 a, this is achieved through a dual-objective mechanism where the generator guides two classification networks to focus on their respective views - global manipulation patterns and local forgery artifacts, while the classifiers' feedback helps the generator maintain proper decomposition granularity. As illustrated in Fig.2 b, our decomposition successfully produces a sparse residual image $I - G(I)$ with a high sparsity ratio (proportion of unused gray levels out of 256 total levels) and significantly lower entropy, containing primarily high-frequency details such as edges and artifacts. This decomposition serves as the foundation for the subsequent cross-view distillation process.

Beyond our main framework, we observe that the decomposition capability learned by our model can be valuable for enhancing existing deepfake detection methods. Since our generator has been trained to effectively separate local forgery artifacts from global facial content, we propose to leverage this pre-trained decomposition model to benefit existing detectors through a lightweight enhancement approach. Specifically, we utilize our pre-trained generator to extract local detail images from input faces and introduce a feature alignment mechanism that encourages existing models to capture these local forgery cues. By enforcing consistency between the local detail features and the original classification features, we enable effective knowledge transfer while maintaining the original model architectures intact. Unlike conventional approaches that rely on explicit feature fusion methods (e.g., feature addition or concatenation) which require architectural modifications, our enhancement method provides a practical solution for upgrading existing detection systems by incorporating local artifact information without structural changes.

The primary contributions of this research can be summarized as follows:

- We propose a novel image decomposition framework that separates facial images into global content and local detail components through a lightweight U-Net generator, enabling separate analysis of different forgery patterns.
- We design a cross-view distillation mechanism to facilitate bidirectional knowledge transfer between global and local views, leading to more comprehensive forgery feature representations.
- We introduce a lightweight enhancement method with feature alignment loss that can improve existing deepfake

detectors without architectural modifications by leveraging decomposed local features.

II. Related Work

In this section, we provide a brief overview of deepfakes and generalizable deepfake detection approaches.

A. Deepfake Forgery

Recent advancements in face manipulation techniques can be broadly categorized into four main areas: face swapping, face reenactment, face attribute editing, and entire image generation through diffusion models.

Face swapping involves replacing a target face in an image or video with a source face while preserving the original attributes [26], [27]. Early approaches relied on subject-specific methods using a shared encoder and individual decoders, requiring separate training for each identity and thus lacking flexibility. A significant breakthrough came with FSGAN [28], [29], which introduced the first subject-agnostic face-swapping method capable of handling both pose and expression variations simultaneously. Following this, FaceShifter [30], [31] achieved improved results through comprehensive integration of target attributes. Recent identity preservation-based methods have further advanced the field by explicitly decoupling identity and attribute features before reconstruction, leading to more natural and convincing results.

Face reenactment, also known as facial expression manipulation, focuses on modifying facial expressions while maintaining identity. Notable approaches include Face2Face [32], [33], which utilizes 3D facial modeling for expression transfer, and NeuralTextures [34], which combines neural textures with rendering networks. An emerging sub-category is talking face generation, which aims to synthesize realistic facial animations synchronized with speech. For instance, HyperReenact [35] demonstrates one-shot reenactment capabilities through joint learning of facial refinement and retargeting.

Face attribute editing enables fine-grained manipulation of specific facial features (e.g., hair color, accessories) while preserving identity. Modern techniques leverage deep learning architectures to achieve selective attribute modification with minimal impact on other facial characteristics [36], [37].

Recently, diffusion models have emerged as a powerful paradigm for entire image generation, with state-of-the-art models such as Stable-Diffusion-2.1 [5], PixArt-α [38], and SiT [39] demonstrating remarkable capabilities in generating high-resolution, photorealistic images from text descriptions using advanced techniques like latent diffusion and self-supervised vision transformers. Further advancing generative model capabilities, StyleShot [40] introduces a framework for generalized style transfer, enabling image creation in diverse styles from text or image inputs. In the domain of character and facial animation, FaceShot [41] proposes a training-free approach to animate varied characters by precisely capturing and retargeting facial expressions from driving videos, contributing to the realism of manipulated video content.

B. Generalizable Deepfake Detection

Recent research efforts have focused on improving the generalization capabilities of deepfake detection through three primary approaches: data augmentation, frequency domain analysis, and identity-aware learning.

Data augmentation approaches aim to simulate diverse forgery artifacts to enhance model robustness. FWA [14] introduces a self-blending strategy that applies transformations to facial regions before warping them back, effectively simulating distortion artifacts inherent in the deepfake generation process. Face X-ray [15] focuses on learning blending boundaries, a common artifact in manipulated images, to improve detection generalization.

Frequency-based methods exploit the distinctive frequency patterns of manipulated images, partially addressing the need for local artifact detection. Qian et al. [16] developed an adaptive frequency-aware division (FAD) approach using learnable filters to capture local frequency statistics through sliding window discrete cosine transform (SWDCT). However, this approach still processes the entire image holistically, potentially limiting its ability to focus on subtle local manipulation traces.

Recent studies have highlighted the critical role of identity information in detection generalization, while also revealing the importance of local feature analysis. Dong et al. [13] identified that binary-labeled training causes models to become overly sensitive to global identity features while potentially overlooking local manipulation artifacts. Their proposed identity-unaware detection framework partially addresses this issue by incorporating multi-scale analysis, though still lacking explicit mechanisms for local artifact detection.

While these approaches have made significant progress in deepfake detection, our work takes a complementary perspective by explicitly separating and analyzing both global and local features. Unlike traditional whole-image processing methods, IDCNet's image decomposition framework enables dedicated learning of local forgery traces while maintaining awareness of global manipulation patterns.

III. The Proposed Method

In this section, we present the technical details of our proposed IDCNet. We begin by formally defining the problem in Section III-A. Section III-B then provides a comprehensive architectural overview of the IDCNet framework. Subsequent subsections will elaborate on the: image decomposition process (Section III-C), the cross-view distillation mechanism (Section III-D), the training and inference procedures (Section III-E), and finally, our strategy for integrating IDCNet's principles with existing detectors (Section III-F).

A. Problem Statement

Given an input facial image $I \in R^{H \times W \times 3}$, the deepfake detection task aims to determine whether I is genuine or manipulated. More challengingly, we seek to develop a detector that can generalize across different manipulation methods $M = \{M_1, M_2, \ldots, M_k\}$ and remain robust against potential perturbations $P(I)$. Formally, we aim to learn a mapping function f:

$$f : I \rightarrow y \in 0, 1, \tag{1}$$

such that:

$$\forall M_i \in M : f(M_i(I)) = 1, \tag{2}$$

$$\forall P : f(I + P(I)) = f(I), \tag{3}$$

where $y = 0$ indicates a real image and $y = 1$ indicates a manipulated image. $M_i(I)$ denotes an image manipulated by the $i - th$ manipulation method, and $P(I)$ represents possible perturbations applied to the input image.

However, relying on a direct mapping $f(I)$ with standard CNNs often proves insufficient due to their inherent tendencies, leading to two key constraints:

1) Global Feature Bias: The model tends to focus on global patterns while overlooking local artifacts. Formally, given a local region $R_i \subset I$ containing manipulation traces:

$$P(y = 1|I) \gg P(y = 1|R_i), \tag{4}$$

indicating the model's prediction is dominated by global features rather than local manipulation evidence.

2) Local Detail Loss: Fine-grained manipulation traces T are often lost during feature extraction. For a manipulation trace $t_i \in T$:

$$||F(I) - F(I \setminus t_i)||_2 \approx 0, \tag{5}$$

where $F(\cdot)$ represents the extracted features, showing the model's insensitivity to local artifacts.

To address these limitations, we propose to decompose the detection into two complementary views:

$$I = I_c + I_r, \tag{6}$$

where $I_c \in R^{H \times W \times 3}$ represents the global content image preserving overall structure and semantic information, while $I_r \in R^{H \times W \times 3}$ captures local details and manipulation artifacts.

The final prediction is derived through:

$$y = h(g_c(I_c), g_r(I_r)), \tag{7}$$

where g_c and g_r are view-specific encoders that extract features from global content and local detail components respectively, and h is a fusion function that combines complementary information from both views.

To ensure effective capture of manipulation traces, we design the decomposition and detection process such that:

$$P(y = 1|I_r) \approx P(y = 1|T), \tag{8}$$

where T represents the set of manipulation traces in the image. This constraint ensures that the detection probability based on local details (I_r) closely matches the probability based on actual manipulation traces (T), thereby maintaining sensitivity to fine-grained forgery artifacts.

B. Architecture Overview

Our IDCNet framework addresses the limitations of existing whole-image processing methods through two key stages: image-level decomposition and cross-view distillation (Fig.3). In the first stage, we employ a lightweight U-Net generator to decompose the input facial image into two complementary components: a global content image I_c that preserves the original image's color and structure, and a local detail image I_r that captures high-frequency artifacts and edge information. This decomposition strategy enables separate analysis of global manipulation patterns and local forgery traces, overcoming the attention bias towards global features in existing approaches.

In the second stage, we treat I_c and I_r as independent views and introduce a novel cross-view distillation mechanism to enable bidirectional knowledge transfer. Through Variational Mutual Distillation (VMD) and Variational Cross Distillation (VCD), each view can benefit from the other's unique perspective - the global view captures overall manipulation patterns while the local view focuses on fine-grained forgery artifacts. This dual-view learning approach enables the model to distill discriminative features from both views, resulting in more comprehensive and robust forgery representations.

C. Image Decomposition

Inspired by image decomposition techniques in computer vision tasks [22], [23], [24], [25], we adopt a U-Net-based generator architecture for our specific deepfake detection scenario. While traditional image-level decoupling approaches often utilize autoencoders [25], which tend to lose critical high-frequency details during encoding-decoding, our U-Net design offers a more suitable solution. Specifically, U-Net's distinctive skip connections enable comprehensive preservation of fine-grained features throughout the network, making it particularly effective for generating residual images that capture subtle forgery artifacts.

As illustrated in Fig.4, our generator architecture employs an initial DoubleConv module that projects the input image into a 64-channel feature space while preserving spatial dimensions. The network follows a symmetric encoder-decoder structure with four Down modules in the encoder path and four corresponding Up modules in the decoder path, connected by skip connections between corresponding layers. Each Down module performs progressive feature extraction by halving spatial dimensions while doubling channel depth, while Up modules restore spatial resolution through transposed convolutions. The skip connections ensure preservation of fine-grained spatial information across the network.

The fundamental DoubleConv building block consists of two sequential 3×3 convolution layers, each followed by batch normalization and ReLU activation. This module is utilized in both Down paths (combined with max-pooling) and Up paths (following transposed convolutions) to maintain consistent feature processing throughout the network.

The reconstruction objective is formulated as:

$$L_{rec} = \text{MSE}(I, I_c), \tag{9}$$

where I denotes the input image and I_c represents the generator output. The subsequent Cross-view Distillation process

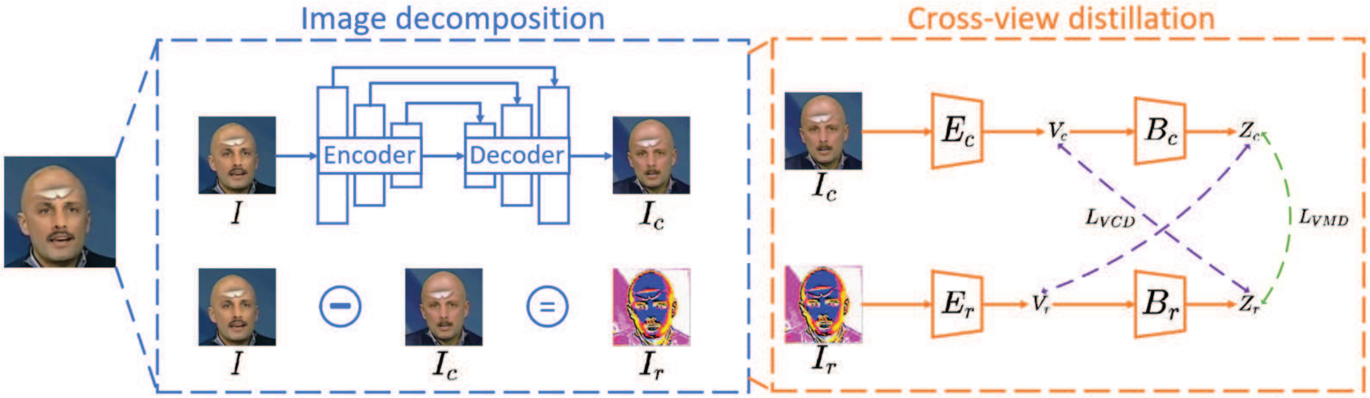


Fig. 3. Overview of the proposed IDCNet framework for deepfake detection. IDCNet framework overview. The model decomposes input facial images into global content (I_c) and local detail (I_r) views. Bidirectional knowledge transfer via Variational Cross Distillation (L_{VCD}) and Variational Mutual Distillation (L_{VMD}) enables effective joint learning of manipulation patterns and forgery artifacts.

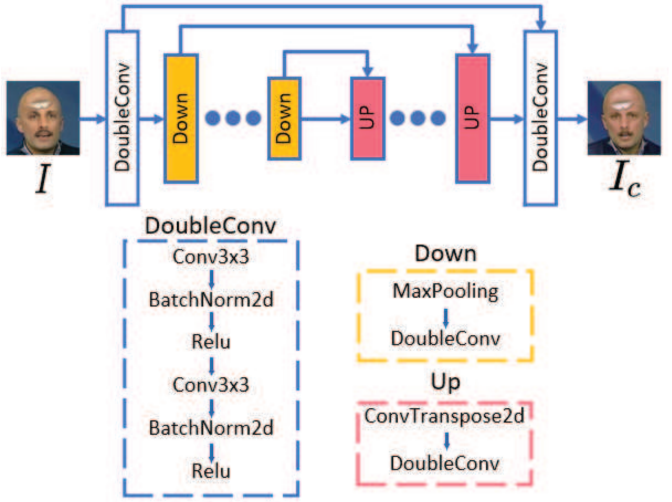


Fig. 4. Architecture of the lightweight U-Net generator used for image-level decomposition, which is designed to precisely control the reconstruction capability to separate global content and local detail components.

ensures preservation of task-relevant features in both content image I_c and residual image I_r.

D. Cross-View Distillation

We treat I_c and I_r as complementary views of the input image, employing separate encoders E_c and E_r to extract global and local forgery features respectively (see Fig.3). Following multi-view learning principles, our goal is to learn representations that capture comprehensive forgery patterns while eliminating task-irrelevant information.

Given observations V_c and V_r (containing information equivalent to I_c and I_r), we aim to learn optimal representations Z_c and Z_r. Using information theory, we decompose the mutual information between V_c and Z_c as:

$$I(V_c; Z_c) = I(V_c; Z_c|V_r) + I(V_r; Z_c), \tag{10}$$

where $I(V_c; Z_c|V_r)$ represents view-specific information in Z_c that cannot be inferred from V_r, and $I(V_r; Z_c)$ denotes view-consistent information shared between Z_c and V_r. Further decomposing $I(V_r; Z_c)$:

$$I(V_r; Z_c) = I(V_r; Z_c|y) + I(V_r; Z_c; y), \tag{11}$$

where $I(V_r; Z_c|y)$ represents task-irrelevant information in Z_c, and $I(V_r; Z_c; y)$ denotes predictive information shared between Z_c, V_r, and target task y. Combining Eqn.(10) and Eqn.(11):

$$I(V_c; Z_c) = I(V_c; Z_c|V_r) + I(V_r; Z_c|y) + I(V_r; Z_c; y). \tag{12}$$

To optimize these representations, we maximize predictive information while minimizing both view-specific and task-irrelevant information through Variational Cross Distillation (VCD) and Variational Mutual Distillation (VMD). [42]:

$$L_{VMD} = D_{KL}[\mathbb{P}_{Z_c} \| \mathbb{P}_{Z_r}], \tag{13}$$

$$L_{VCD} = D_{KL}[\mathbb{P}_{V_r} \| \mathbb{P}_{Z_c}], \tag{14}$$

where $D_{KL}[\cdot\|\cdot]$ denotes the KL divergence between predictive distributions $\mathbb{P}_{(\cdot)}$. Specifically, when the predictive distributions of two views are aligned (i.e., $D_{KL}[\mathbb{P}_{Z_c} \| \mathbb{P}_{Z_r}] = 0$), it implies that they contain consistent information about the target task, thereby eliminating information specific to only one of the views. Minimizing $D_{KL}[\mathbb{P}_{V_r} \| \mathbb{P}_{Z_c}]$ forces Z_c to accurately predict the information about the target task contained in V_r, thereby discarding any task-irrelevant information. In other words, if Z_c can accurately predict the target labels expected from V_r, any information within Z_c that is not related to the target labels is treated as noise and will be eliminated.

The complete cross-view distillation loss, applied symmetrically to both views, is defined as:

$$\begin{aligned} L_{cvd} = \underbrace{D_{KL}[\mathbb{P}_{Z_c} \| \mathbb{P}_{Z_r}] + D_{KL}[\mathbb{P}_{Z_r} \| \mathbb{P}_{Z_c}]}_{L_{VMD}} \\ + \underbrace{D_{KL}[\mathbb{P}_{V_r} \| \mathbb{P}_{Z_c}] + D_{KL}[\mathbb{P}_{V_c} \| \mathbb{P}_{Z_r}]}_{L_{VCD}}. \end{aligned} \tag{15}$$

This loss function encourages maximizing the relevant information from both views while minimizing residual content and noise introduced by incomplete decoupling.

For feature extraction, we employ established architectures like XceptionNet [43] and EfficientNet [44] as encoder

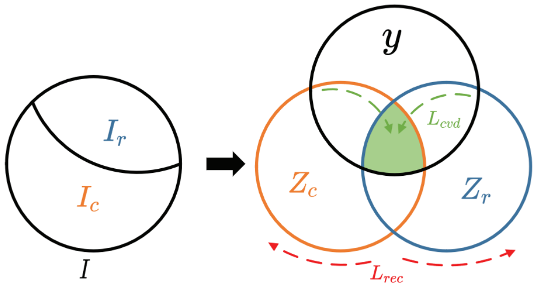


Fig. 5. Venn diagram illustrating the mutual information optimization process. Through L_{rec}, input image I is decomposed into I_c and I_r, yielding increasingly distinct representations Z_c and Z_r. The L_{cvd} loss enhances shared task-relevant information (green area) between these representations while maintaining their complementarity.

backbones. The bottleneck modules B_c and B_r compress feature dimensionality through a two-stage process: (1) initial dimensionality reduction via linear projection with batch normalization and non-linear activation, and (2) final mapping to target dimensions. This architecture preserves discriminative features while facilitating efficient cross-view alignment.

E. Training and Inference

Our framework integrates Image Decoupling and Cross-view Distillation through a joint optimization objective:

$$L_{total} = \alpha L_{rec} + L_{cvd}, \tag{16}$$

where α is a balancing coefficient that controls the trade-off between reconstruction loss (L_{rec}) and cross-view distillation loss (L_{cvd}). As illustrated in Fig.5, the Image Decoupling module first decomposes input I into content component I_c and residual component I_r through reconstruction learning. The Cross-view Distillation then optimizes their latent representations Z_c and Z_r by leveraging L_{cvd} to maintain task-relevant shared information while the reconstruction process naturally encourages their distinctiveness. A properly tuned α ensures an optimal balance between component decomposition and shared information preservation, enabling the model to capture complementary features from different perspectives.

During inference, we utilize Z_r for final prediction to avoid potential overfitting to content features.

F. Integration With Existing Detectors

Our framework can be seamlessly integrated into existing detectors through a lightweight feature alignment loss to enhance their sensitivity to local forgery artifacts. As illustrated in Fig.6, the integration process employs the pre-trained image decoupling model to extract the local detail component I_r from the input image, followed by applying a consistency constraint between feature representations of the original and local detail images. The enhancement is achieved through the feature alignment loss:

$$L_{FA} = \mathrm{MSE}(F, F_r), \tag{17}$$

where F and F_r denote the feature representations from the original and local detail images respectively.

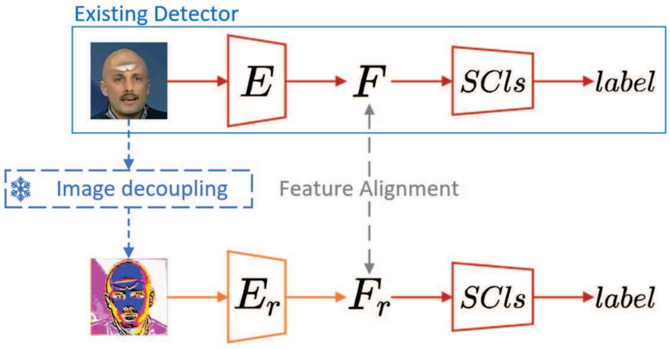


Fig. 6. Integration framework for enhancing existing detectors. Our pre-trained Image decoupling model extracts complementary local detail features from input images. $SCls$ denotes the shared classifier. The alignment between global features F and local forgery features F_r enables seamless enhancement of existing detection models without architectural modifications.

IV. Experiments

This section presents comprehensive experimental evaluations. First, we assess the generalization capability of IDCNet by comparing it with nine baseline methods from DeepfakeBench [45] across seven datasets, covering 14 different manipulation techniques. Second, we conduct ablation studies to validate the effectiveness of each component in IDCNet. Then, we verify the performance gains brought by our feature alignment loss on four baseline models. Finally, we perform robustness tests to evaluate the model's performance under various image perturbations.

A. Experimental Settings

1) Datasets: We conduct comprehensive experiments on seven representative deepfake detection datasets: FaceForensics++ (FF++) [10], CelebDF-v1 (CDFv1) [59], CelebDF-v2 (CDFv2) [59], DeepFake Detection (DFD),[1] DeepFake Detection Challenge (DFDC) [60], DeepFake Detection Challenge Preview (DFDCP) [61], and DF40 [53].

The FF++ dataset comprises 1,000 pristine videos and their corresponding 4,000 manipulated versions generated using four distinct manipulation techniques: DeepFakes (DF),[2] Face2Face (F2F) [32], FaceSwap (FS) [62], and NeuralTextures (NT) [34]. The CelebDF dataset, available in two versions, demonstrates progressive improvements in synthesis quality. CDFv1 contains 408 real videos and 795 DeepFake-generated videos with enhanced visual quality through an improved synthesis algorithm. Its successor, CDFv2, expands the collection to 590 real videos and 5,639 synthetic videos, featuring substantial improvements in visual fidelity.

For broader evaluation scope, we incorporate DFD, which consists of 363 real videos and 3,068 manipulated videos generated through various synthesis techniques. The DFDC dataset provides a large-scale benchmark with over 100,000 video clips featuring 3,426 paid actors, synthesized using diverse DeepFake and GAN-based approaches. DFDCP, containing approximately 5,000 videos, serves as a preliminary

[1] https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[2] https://github.com/deepfakes/faceswap

TABLE I

Cross-Dataset Generalization Performance. All Models Are Trained on FF++(c23) [10] and Evaluated on Six Datasets. FF++ Represents in-Dataset Evaluation Results, While the Other Five Datasets Demonstrate Cross-Dataset Generalization Capability. The Best and Second-Best Results Are Highlighted in Bold and Underlined, Respectively. All Baseline Models Are Implemented Based on the DeepFake Detection Benchmark [45]

Methods	FF++		CDFv1		CDFv2		DFD		DFDC		DFDCP	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
SRM[17]	0.6201	0.9076	0.5562	0.6899	0.6307	0.6946	0.6261	0.7676	0.4339	0.6162	0.4626	0.6777
CNN-Aug[46]	0.8374	0.7553	0.7148	0.6526	0.6751	0.6318	0.633	0.5947	0.5649	0.5511	0.5522	0.5426
Capsule[47]	0.8851	0.8062	0.7688	0.7026	0.7657	0.6988	0.7059	0.6471	0.6021	0.5711	0.5453	0.5379
FFD[48]	0.9588	0.8987	0.7058	0.6566	0.6871	0.6386	0.7915	0.7143	0.6573	0.6056	0.6754	0.6276
CORE[49]	0.9706	0.9163	0.7174	0.6585	0.7417	0.679	0.8266	0.7464	0.6566	0.6061	0.6806	0.635
SPSL[50]	0.9453	0.8709	0.8028	0.7389	0.7323	0.6656	0.8202	0.7394	0.6541	0.6043	0.6912	0.6402
Recce[51]	0.9738	0.9205	0.7316	0.6712	0.7421	0.6797	0.8448	0.7632	0.6584	0.6058	0.7046	0.6424
F3Net[16]	0.9782	0.9286	0.7521	0.6823	0.7306	0.6687	0.7916	0.7152	0.6765	0.6174	0.7339	0.6668
UCF[52]	**0.9812**	**0.9364**	0.8118	0.7271	0.7731	0.6977	0.8209	0.7424	0.6838	0.6208	0.6854	0.6257
Ours	0.9696	0.9187	**0.8144**	**0.7992**	**0.8089**	**0.7361**	**0.8471**	**0.7771**	**0.7244**	**0.6605**	**0.7409**	**0.6869**

evaluation platform while ensuring demographic diversity across gender, skin tone, and age distributions.

To evaluate the model's generalization capability across contemporary generation techniques, we further conduct experiments on recent GAN- and diffusion-based manipulations. Specifically, we select seven representative methods from the DF40 dataset: FSGAN [28], InSwapper,[3] BlendFace [26], FS-Vid2Vid [63], HyperReenact [35], Stable Diffusion (SD2.1) [5], PixelArt-α [38], and SiT [39].

2) Implementation Details: For data preprocessing, we employ the DLIB library for face detection and alignment. Following [45], we set the margin ratio to 1.3 for face cropping and resize the cropped facial images to 256 × 256 pixels. For both training and testing, we sample 32 frames from each video. The framework is implemented in PyTorch [64] and trained on two NVIDIA A5000 GPUs. We utilize ResNet50 [57] and EfficientNetB4 [44] architectures pre-trained on ImageNet [65] as the content encoder E_c and local detail encoder E_r, respectively. The model is optimized using Adam [95] with an initial learning rate of 1×10^{-4} and a batch size of 16. The balancing coefficient α is empirically set to 1. All other models participating in the comparison utilized the pretrained weights available in DeepfakeBench [45].

3) Evaluation Metrics: Following [66], we employ two metrics for comprehensive performance evaluation: Area Under the receiver operating characteristic Curve (AUC) and Equal Error Rate (EER). The AUC metric quantifies the model's overall discrimination ability across different classification thresholds, while EER represents the operating point where false acceptance rate equals false rejection rate, providing a balanced assessment of detection performance.

4) Baselines: To ensure fair comparison in generalization evaluation, all nine baseline methods are sourced from DeepfakeBench [45], following identical experimental settings and evaluation protocols. Our IDCNet adheres to the same experimental configuration.

To evaluate the effectiveness of the feature alignment strategy, we reimplemented four classic deepfake detection methods and applied our feature alignment approach. This choice allows us to focus on evaluating the impact of our feature alignment strategy while minimizing potential implementation uncertainties that might arise from complex architectures. These methods are different from the nine baselines used in the generalization evaluation, though they follow the same experimental settings as specified in DeepfakeBench [45].

B. Generalization Evaluation

The experimental results reveal a critical challenge in deepfake detection: the substantial performance degradation when models encounter unseen manipulation techniques or datasets. As shown in Table I, while state-of-the-art methods like UCF achieve exceptional performance on FF++ (AUC of 0.9812 and EER of 0.9364), they exhibit significant performance drops in cross-dataset scenarios. This performance gap highlights the limitations of existing approaches in handling domain shifts and diverse manipulation techniques.

Our proposed method addresses this challenge effectively by maintaining competitive in-dataset performance (AUC of 0.9696 and EER of 0.9187) while demonstrating superior generalization capability across all five cross-dataset scenarios. The performance improvements are particularly pronounced on challenging datasets: on CDFv2, our method achieves gains of 4.4% in AUC (0.8089 vs 0.7731) and 5% in EER (0.7361 vs 0.6988) compared to the second-best method. Even more significant improvements are observed on the DFDC dataset, with increases of 5.6% in AUC (0.7244 vs 0.6838) and 6% in EER (0.6605 vs 0.6208). These consistent improvements across different datasets validate the effectiveness of our architectural design in capturing generalizable manipulation artifacts.

To further evaluate our method's robustness against emerging manipulation techniques, we conduct extensive experiments on the recent DF40 dataset [53]. As shown in Table II, our method demonstrates remarkable generalization capability by consistently outperforming existing approaches across diverse generation types. Specifically, our method achieves superior performance in detecting GAN-based face swapping (improvements of up to 5.5% AUC on InSwapper), video reenactment (improvements of 5.7% AUC on HyperReEnact), and diffusion model-based generation (improvements of

[3] https://github.com/haofanwang/inswapper

TABLE II

GENERALIZATION PERFORMANCE ON ADVANCED GENERATION METHODS. DETECTION RESULTS (AUC) ON THE DF40 DATASET [53], WHICH CONTAINS DIVERSE MANIPULATION TECHNIQUES INCLUDING GAN-BASED FACE SWAPPING (FSGAN, INSWAPPER, BLENDFACE), VIDEO REENACTMENT (FSVID2VID, HYPERREENACT), DIFFUSION MODELS (SD2.1, SIT), AND ARTISTIC STYLE TRANSFER (PIXELART). THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINED, RESPECTIVELY

Methods	DF40							
	FSGAN	**InSwapper**	**BlendFace**	**FSVid2Vid**	**HyperReEnact**	**SD2.1**	**PixelArt**	**SiT**
FFD[48]	0.8421	0.6265	0.5944	0.6262	0.5601	0.8922	0.7253	0.5491
CNN-Aug[46]	0.6933	0.6527	0.6654	0.6790	0.6227	0.8304	0.7148	0.5992
SPSL[50]	0.7778	0.6058	0.639	0.6834	0.7211	0.7487	0.7234	0.5801
Capsule[47]	0.7419	0.7112	0.6023	0.6805	0.6399	0.8819	0.6495	0.6102
SRM[17]	0.8162	0.7485	0.6812	0.7071	0.6567	0.6151	0.7975	0.5868
F3Net[16]	0.8567	0.6941	0.7256	<u>0.7276</u>	0.6614	0.8838	0.7975	0.5961
UCF[52]	<u>0.8628</u>	0.7124	0.701	0.7255	0.6622	<u>0.9252</u>	0.836	0.6092
CORE[49]	0.8582	0.7457	0.7045	0.7075	<u>0.7263</u>	0.7310	0.8242	**0.6512**
Recce[51]	0.8408	<u>0.7719</u>	<u>0.7294</u>	0.7099	0.7252	0.7835	<u>0.8507</u>	0.6302
Ours	**0.8661**	**0.8147**	**0.7371**	**0.7370**	**0.7707**	**0.9295**	**0.8737**	<u>0.6311</u>

TABLE III

EFFECTIVENESS OF THE PROPOSED DISTILLATION STRATEGY. THE TABLE PRESENTS CROSS-DATASET EVALUATION RESULTS OF FOUR REPRESENTATIVE DETECTORS BEFORE AND AFTER APPLYING OUR FEATURE ALIGNMENT (FA) STRATEGY. ALL MODELS ARE TRAINED ON FF++(C23) [10] AND EVALUATED ON FIVE CHALLENGING DATASETS. PERFORMANCE CHANGES (↑ / ↓) ARE CALCULATED RELATIVE TO THE BASELINE METHODS. THE BEST RESULTS FOR EACH DETECTOR PAIR ARE SHOWN IN BOLD

Model	**CDFv1**		**CDFv2**		**DFD**		**DFDC**		**DFDCP**	
	AUC	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**
GramNet[54]	0.5272	0.5094	0.6418	0.6036	0.8022	0.7371	0.6894	0.6312	0.732	0.6743
GramNet-FA	**0.6938** **(↑24.0%)**	**0.6481** **(↑21.4%)**	**0.7232** **(↑11.3%)**	**0.6606** **(↑8.6%)**	**0.8669** **(↑7.5%)**	**0.7827** **(↑5.8%)**	**0.7007** **(↑1.6%)**	**0.6347** **(↑0.6%)**	**0.7509** **(↑2.5%)**	**0.6793** **(↑0.7%)**
MultiAttention[55]	0.6351	0.5872	0.7086	0.6568	0.7972	0.7205	0.7155	0.654	**0.7903**	**0.7135**
MultiAttention-FA	**0.7586** **(↑16.3%)**	**0.6999** **(↑16.1%)**	**0.7608** **(↑6.9%)**	**0.6925** **(↑5.2%)**	**0.8085** **(↑1.4%)**	**0.7307** **(↑1.4%)**	**0.7394** **(↑3.2%)**	**0.6684** **(↑2.2%)**	0.7557 (↓4.4%)	0.6926 (↓2.9%)
TwoStreamNet[17]	0.6001	0.5721	0.6712	0.6277	**0.8471**	**0.7674**	0.7095	0.6457	**0.7639**	0.6948
TwoStreamNet-FA	**0.7262** **(↑17.4%)**	**0.6686** **(↑14.4%)**	**0.7472** **(↑10.2%)**	**0.6805** **(↑7.8%)**	0.7906 (↓6.7%)	0.722 (↓5.9%)	**0.7264** **(↑2.3%)**	**0.6593** **(↑2.1%)**	0.7635 (↓0.1%)	**0.7022** **(↑1.1%)**
SIA[56]	0.5914	0.5593	0.6933	0.6411	**0.8498**	**0.7735**	0.7021	0.6387	0.7526	0.6871
SIA-FA	**0.7961** **(↑25.7%)**	**0.7103** **(↑21.3%)**	**0.8079** **(↑14.2%)**	**0.7291** **(↑12.1%)**	0.8221 (↓3.3%)	0.7511 (↓2.9%)	**0.7529** **(↑6.7%)**	**0.6832** **(↑6.5%)**	**0.7916** **(↑4.9%)**	**0.7211** **(↑4.7%)**

0.4% AUC on SD2.1). The consistent strong performance across both traditional and emerging manipulation techniques convincingly demonstrates our method's robust generalization ability.

C. Integration With Existing Detectors

To validate the effectiveness and applicability of our proposed feature alignment strategy, we conduct extensive experiments on four representative deepfake detection models: GramNet [54], MultiAttention [55], TwoStreamNet [17], and SIA [56]. Each model is first trained using its original configuration on FF++(c23) to establish baseline performance, then enhanced with our proposed feature alignment loss.

The experimental results presented in Table III demonstrate that our feature alignment strategy substantially enhances the cross-dataset generalization capability of existing detectors. The improvements are particularly pronounced when dealing with significant domain shifts, as evidenced by the results on CDFv1 and CDFv2 datasets. The SIA model, when enhanced with our feature alignment strategy (SIA-FA), achieves remarkable improvements with AUC gains of 25.7% on CDFv1 (improving from 0.5914 to 0.7961) and 14.2% on CDFv2 (increasing from 0.6933 to 0.8079). Similarly, GramNet-FA demonstrates substantial enhancements with AUC improvements of 24.0% and 11.3% on these datasets, respectively. These significant gains suggest that our feature alignment strategy effectively bridges the domain gap between training and testing distributions.

The performance improvements extend to more recent and challenging datasets as well. On the DFDC dataset, which features more sophisticated manipulation techniques, all enhanced models show consistent improvements. SIA-FA leads the improvements with an AUC gain of 6.7% and an EER improvement of 6.5%, while other enhanced models demonstrate moderate but consistent gains. The DFDCP dataset results follow a similar pattern, with SIA-FA achieving notable improvements of 4.9% in AUC and 4.7% in EER, demonstrating the strategy's effectiveness across different manipulation techniques.

Interestingly, the results on the DFD dataset reveal nuanced patterns in the effectiveness of feature alignment. While GramNet-FA and MultiAttention-FA show positive gains with AUC improvements of 7.5% and 1.4% respectively, TwoStreamNet-FA and SIA-FA experience slight performance degradation (3.3-6.7% decrease in AUC). This variation in performance suggests that the effectiveness of feature alignment

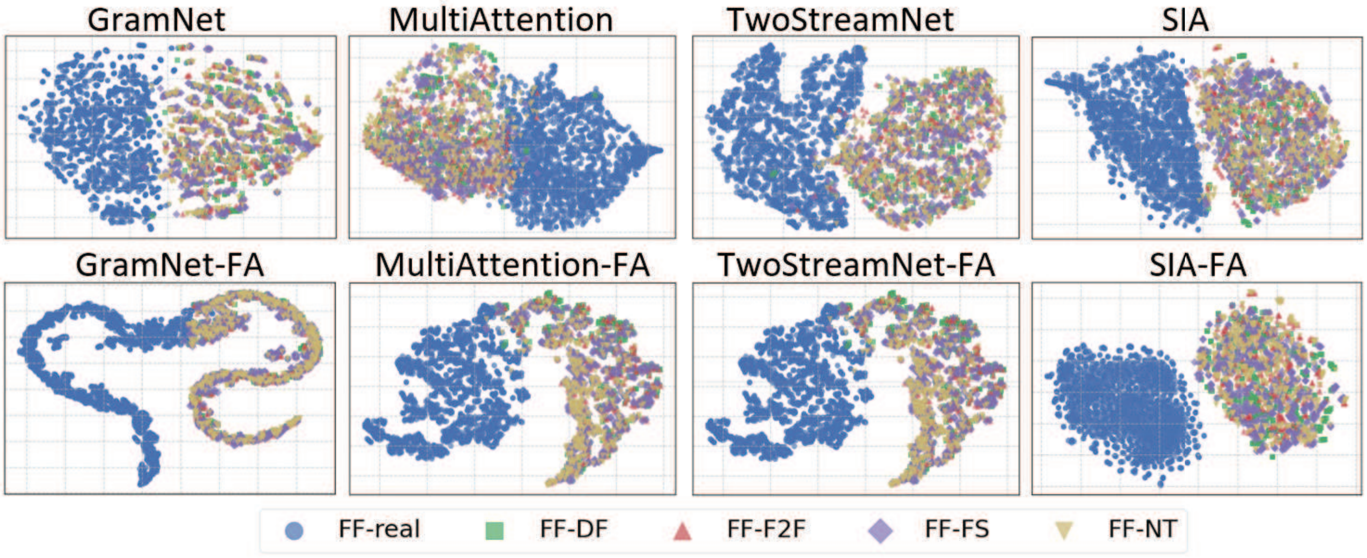


Fig. 7. T-SNE visualization comparing feature distributions of four representative detectors before and after the proposed Feature Alignment (FA) strategy. The visualization demonstrates the enhanced feature separability between real and fake samples achieved through our approach.

may be influenced by the inherent architectural characteristics of the base models and their interaction with specific domain shift patterns present in the DFD dataset. For instance, for models like TwoStreamNet or SIA that might rely on specialized cues (e.g., frequency domain or specific inconsistency patterns), the alignment with our decomposed spatial local features might lead to a nuanced trade-off on datasets where those original cues are particularly dominant.

We employ t-SNE [67] to visualize the feature distributions of the aforementioned four models before and after applying our feature alignment strategy. As illustrated in Fig.7, the feature alignment strategy significantly enhances the separability between real samples and samples from four different types of forgeries, demonstrating the effectiveness of our approach in learning more discriminative features for deepfake detection.

D. Ablation Study

1) Various Feature Extractors: To thoroughly investigate the impact of encoder architectures on model generalization, we conduct extensive experiments with different backbone combinations for the content encoder (E_c) and local detail encoder (E_r). Four representative architectures are evaluated: ResNet50 [57], Xception [43], EfficientNetB4 [44], and SFIResNet [58]. Among these, SFIResNet is specifically designed for deepfake detection, utilizing spatial-frequency interactive convolution to construct a backbone network that better captures forgery traces.

The experimental results in Table IV reveal several important insights about the architectural choices for our dual-encoder framework. When SFIResNet is employed as the local detail encoder (E_r), the model's performance is consistently suboptimal across different content encoder configurations. This pattern is particularly evident in cross-dataset scenarios, where combinations with SFIResNet as E_r achieve relatively low AUC scores (ranging from 0.61 to 0.70) on challenging datasets like DFDC and DFDCP. This performance limitation suggests that SFIResNet's frequency-aware architecture, while theoretically promising for forgery detection, may not be optimal for extracting fine-grained local manipulation patterns.

Conversely, the utilization of SFIResNet as the content encoder (E_c) yields notably different results. The combination of SFIResNet as E_c and EfficientNetB4 as E_r demonstrates exceptional performance, achieving the highest in-domain AUC of 0.9709 on FF++ while maintaining robust cross-domain generalization with AUC scores of 0.8091 on CDFv2 and 0.7242 on DFDC. This superior performance can be attributed to the complementary strengths of these architectures: SFIResNet's spatial-frequency interactive features effectively capture global manipulation patterns, while EfficientNetB4's efficient architecture excels at extracting local detail features.

The ResNet50-EfficientNetB4 combination also demonstrates strong performance, achieving consistent results across different datasets (AUC of 0.9696 on FF++, 0.8144 on CDFv1, and 0.7244 on DFDC). This suggests that the combination of ResNet50's robust feature extraction with EfficientNetB4's efficient architecture provides a reliable foundation for cross-domain generalization.

2) Effects of Variational Mutual Distillation and Variational Cross Distillation: To comprehensively evaluate the effectiveness of Variational Mutual Distillation (VMD) and Variational Cross Distillation (VCD), we conducted extensive ablation studies across multiple datasets. We systematically investigated four variants: baseline without mutual information constraints, VMD-only, VCD-only, and the combined VMD-VCD approach. As demonstrated in Table V, the integration of both VMD and VCD consistently yields superior performance across all evaluation metrics.

Specifically, the combined VMD-VCD strategy achieves the highest AUC scores across most datasets: 0.9696 on FF++, 0.8089 on CDFv2, 0.8471 on DFD, and 0.7244 on DFDC. This consistent improvement is particularly noteworthy in

TABLE IV

Ablation Study on Encoder Architectures. The Table Presents Comprehensive Evaluation Results Using Different Backbone Combinations for Content Encoder (E_c) and Local Detail Encoder (E_r). All Models Are Evaluated on FF++ [10] and Five Cross-Domain Datasets. Four Representative Architectures Are Investigated: ResNet50 [57], Xception [43], EfficientNetB4 [44], and the Frequency-Aware SFIResNet [58]. The Best Results for Each Detector Pair Are Shown in Bold

E_c	E_r	FF++		CDFv1		CDFv2		DFD		DFDC		DFDCP	
		AUC	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**
ResNet50	SFIResNet	0.8155	0.7484	0.8112	0.7356	0.7236	0.6635	0.6016	0.5765	0.6221	0.5881	0.6237	0.5791
SFIResNet	SFIResNet	0.7843	0.7121	0.7334	0.6837	0.7266	0.6721	0.6684	0.6234	0.6211	0.5866	0.6841	0.6184
EfficientNetB4	SFIResNet	0.8439	0.7637	0.7818	0.7026	0.7495	0.6856	0.692	0.6412	0.6385	0.5974	0.7046	0.6331
Xception	SFIResNet	0.9005	0.8187	0.8203	0.7382	0.7474	0.6831	0.7143	0.6598	0.6328	0.5931	0.7006	0.6483
ResNet50	ResNet50	0.9601	0.9002	0.7364	0.6813	0.7626	0.6988	0.7762	0.7124	0.6175	0.5853	0.6803	0.6309
Xception	ResNet50	0.9421	0.8719	0.7614	0.7085	0.7519	0.6898	0.8374	0.7685	0.6143	0.5798	0.6392	0.5939
EfficientNetB4	ResNet50	0.9565	0.8931	0.7468	0.6941	0.7341	0.6768	0.8198	0.7531	0.6345	0.5935	0.7441	0.6754
SFIResNet	ResNet50	0.9684	0.9111	0.7231	0.6516	0.7693	0.6946	0.8011	0.7309	0.6777	0.6295	0.7227	0.6617
Xception	EfficientNetB4	0.9261	0.8473	0.8002	0.7163	0.7897	0.7158	0.8221	0.7492	0.6537	0.6093	0.7067	0.6662
Xception	Xception	0.9643	0.9103	0.7661	0.6993	0.7713	0.7072	0.8405	0.7683	0.6661	0.6164	0.6874	0.6237
ResNet50	Xception	0.9427	0.8778	0.7767	0.7042	0.7881	0.7141	0.8402	0.7647	0.6655	0.6164	**0.7971**	**0.7281**
EfficientNetB4	EfficientNetB4	0.9664	0.9172	0.7759	0.7229	0.7974	0.7285	0.8254	0.7535	0.7046	0.6381	0.7372	0.6621
EfficientNetB4	Xception	0.9601	0.9047	0.7984	0.7203	0.7958	0.7187	**0.8522**	0.7746	0.6727	0.6222	0.7472	0.6795
SFIResNet	Xception	0.9531	0.8877	**0.8215**	0.7484	0.8066	0.7326	0.8395	0.7712	0.6913	0.6364	0.7742	0.7042
SFIResNet	EfficientNetB4	**0.9709**	**0.9254**	0.7737	0.7144	**0.8091**	0.7357	0.8274	0.7737	0.7242	**0.6662**	0.7644	0.6911
ResNet50	EfficientNetB4	0.9696	0.9187	0.8144	**0.7592**	0.8089	**0.7361**	0.8471	**0.7771**	**0.7244**	0.6605	0.7492	0.6869

TABLE V

Ablation Study on the Effectiveness of Variational Mutual Distillation (VMD) and Variational Cross Distillation (VCD) Strategies. Models Are Trained on FF++(c23) and Evaluated Across Six Datasets, With Best Results Shown in Bold

VMD	VCD	FF++		CDFv1		CDFv2		DFD		DFDC	
		AUC	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**
		0.9499	0.8926	0.7126	0.6578	0.7636	0.7038	0.8394	0.7671	0.6779	0.6325
✓		0.9548	0.8938	0.7728	0.7077	0.7946	0.7204	0.8057	0.7391	0.6738	0.6265
	✓	0.9571	0.8950	**0.8171**	0.7575	0.7987	0.7234	0.8331	0.7667	0.7044	0.6511
✓	✓	**0.9696**	**0.9187**	0.8144	**0.7992**	**0.8089**	**0.7361**	**0.8471**	**0.7771**	**0.7244**	**0.6605**

challenging cross-dataset scenarios, where the model demonstrates enhanced generalization capability. The performance gains are further corroborated by corresponding improvements in EER metrics, with the combined approach achieving optimal EER values across all datasets (0.9187 on FF++, 0.7992 on CDFv1, 0.7361 on CDFv2, 0.7771 on DFD, and 0.6605 on DFDC).

These empirical results strongly suggest that the synergistic combination of VMD and VCD effectively enhances the model's feature learning and representation capabilities, leading to more robust and generalizable deepfake detection performance across diverse datasets.

3) Effects of Balancing Coefficient α: We investigate the impact of balancing coefficient α on model performance by varying it from 0 to 1.0, as shown in Table VI. For in-domain testing on FF++, the model achieves the highest AUC of 0.9696 when $\alpha = 1.0$. This setting also yields the best performance on the DFD (0.8471) and DFDC (0.7244) cross-domain datasets. While $\alpha = 0.25$ shows optimal results on CDFv1 (0.8652) and CDFv2 (0.8355), the performance with $\alpha = 1.0$ remains competitive on these datasets (0.8144 and 0.8089 respectively). Considering its leading performance on the training dataset (FF++) and strong results across a majority of the cross-domain evaluations (DFD, DFDC), $\alpha = 1.0$ was selected for our main experiments as it demonstrated a robust overall performance profile. These findings suggest that a significant emphasis on reconstruction learning (controlled by α) is beneficial for learning generalizable robust features, especially when coupled with strong source domain performance.

TABLE VI

Impact of Balancing Coefficient α on Model Performance Across Different Datasets. The Table Demonstrates How Different Trade-Offs Between Reconstruction Loss (L_{rec}) for Component Decomposition and Cross-View Distillation Loss (L_{cvd}) Affect AUC Scores for Both in-Domain (FF++) and Cross-Domain (CDFv1, CDFv2, DFD, DFDC) Evaluations

α	**FF++**	**CDFv1**	**CDFv2**	**DFD**	**DFDC**
0	0.9414	0.7935	0.7671	0.7008	0.6959
0.25	0.9469	**0.8652**	**0.8355**	0.8072	0.6973
0.5	0.9532	0.8637	0.7946	0.8328	0.7241
0.75	0.9476	0.8273	0.8184	0.7881	0.7055
1.0	**0.9696**	0.8144	0.8089	**0.8471**	**0.7244**

E. Robustness Evaluation

Online transmitted deepfake videos inevitably undergo unknown transformations, and various forms of image degradation can eliminate low-level forgery cues [68]. To validate both the robustness of our method and the effectiveness of Feature Alignment (FA), we conduct comprehensive

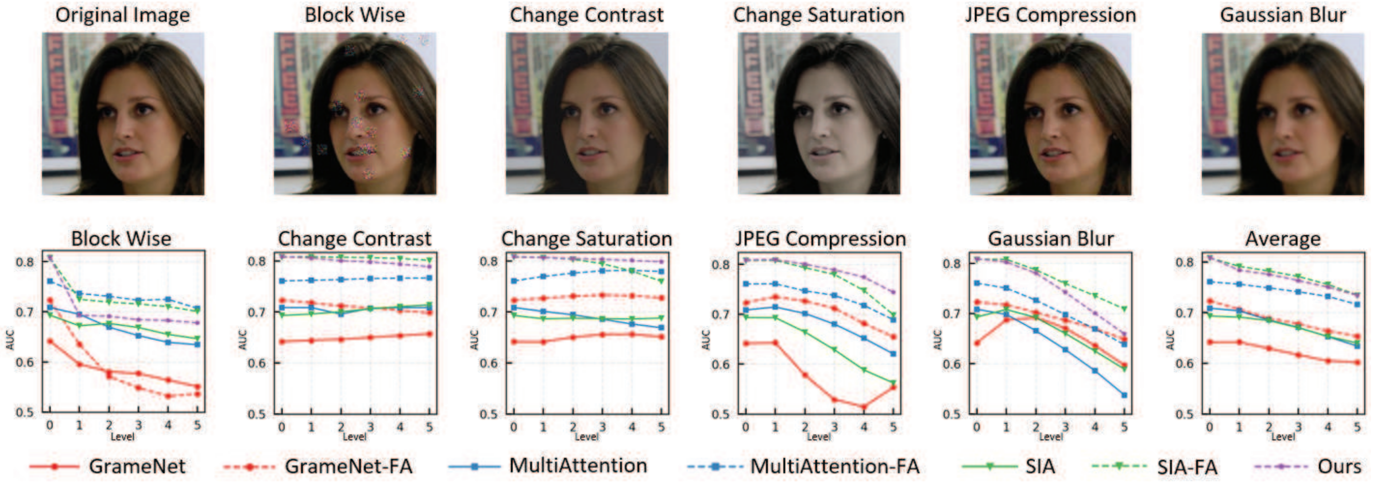


Fig. 8. Robustness evaluation against unseen perturbations. We report AUC scores under five degradation levels of five specific perturbation types [23]. “Average” indicates the mean AUC across all perturbations at each level. Results are shown for our method, four representative baselines, and their variants with feature alignment, demonstrating the robustness of both our approach and the feature alignment strategy.

experiments on CDFv2 dataset following the protocol in [69]. We evaluate five common perturbations across five severity levels (Level-1 to Level-5).

As shown in Fig.8, while our method exhibits strong robustness across various perturbations, Block Wise corruption and Gaussian Blur induce the most pronounced performance decline, particularly at higher severity levels. JPEG compression also affects performance, especially with increasing compression ratios. This vulnerability to structure-altering and high-frequency-attenuating perturbations can be attributed to IDCNet’s core mechanism, which relies on decomposing the image into global content (I_c) and a local detail view (I_r) designed to capture subtle high-frequency artifacts and edge information crucial for forgery detection. Perturbations like Block Wise and Gaussian Blur directly degrade or remove these fine-grained local cues within I_r, thereby diminishing the discriminative information available for the local detail encoder (E_r) and the subsequent prediction, which relies on the local view’s features (Z_r)[cite: 141]. In contrast, color-based perturbations such as changes in contrast and saturation show minimal impact, as they primarily affect the global content view (I_c) while potentially preserving the integrity of the high-frequency local artifacts in I_r.

F. Efficiency Analysis

The U-Net generator, integral to our image decomposition process (Section III-C), operates directly in the pixel space. This architectural choice is fundamental for enabling the precise separation of an input image into its global content and local detail components by allowing direct manipulation and reconstruction. This section provides an analysis of the associated memory and time consumption for our proposed IDCNet framework, including considerations for these pixel-space operations.

IDCNet’s inference efficiency is detailed in Table VII. It achieves competitive latency (8.21 ms/image) and GPU memory usage (1113.12 MiB), performing favorably against several baselines (e.g., Recce, SRM, UCF), though some lighter models are faster. While IDCNet’s GFLOPs (60.14) and parameters (56.78 M) are higher than some methods, this stems from its dual-encoder and U-Net architecture, essential for its advanced decomposition and feature learning capabilities. We consider this increased computational cost a justifiable trade-off for the significant improvements in detection accuracy and generalization detailed in our experimental results (Section IV-B).

TABLE VII

Comparative Inference Efficiency of Deepfake Detection Models

Model Name	Inference Latency (ms/image)	Peak GPU Memory (MiB)	GFLOPs	Params (M)
Recce	94.55	660.02	8.06	23.78
SRM	12.74	1161.16	13.81	53.23
UCF	10.34	1149.18	12.19	44.51
IDCNet	**8.21**	**1113.12**	**60.14**	**56.78**
Capsule	5.96	116.56	16.21	3.89
F3Net	5.48	453.64	6.05	20.81
CORE	5.27	527.05	6.01	20.81

TABLE VIII

Model Efficiency During FA Training. Baseline Rows Show Inference Stats for Reference. ’-FA’ Rows Detail Resource Use When the U-Net Is Active With Baselines During FA Training. The FA Strategy Adds No Inference Overhead to Final Enhanced Models

Model Name	Time (ms/image)	Peak GPU Mem. (MiB)	GFLOPs	Params (M)
SIA	19.23	781.24	0.13	0.15
SIA-FA	39.18	1574.11	55.03	33.14
GramNet	4.89	107.36	2.96	11.73
GramNet-FA	25.04	659.17	57.87	58.85
MultiAttention	17.03	730.48	0.11	0.13
MultiAttention-FA	38.52	1758.36	55.35	34.09
TwoStreamNet	11.97	1074.09	13.81	53.23
TwoStreamNet-FA	32.18	1869.06	68.71	86.76

Table VIII details the computational resources utilized during the Feature Alignment (FA) training phase. When integrating our pre-trained U-Net generator with existing detectors (e.g., SIA, GramNet, MultiAttention, TwoStreamNet) for FA, the U-Net is active to produce the local detail component I_r. This, as expected, increases the training time per image and peak GPU memory consumption compared to training the baseline models alone. For instance, SIA-FA's training time increases to 39.18 ms/image from SIA's 19.23 ms/image, and peak GPU memory rises to 1574.11 MiB from 781.24 MiB. Similar trends are observed for GramNet-FA, MultiAttention-FA, and TwoStreamNet-FA. The GFLOPs and parameter counts for the "-FA" variants also reflect the addition of the U-Net (with approximately 31.03 M parameters and 54.9 GFLOPs, as derived from the difference between SIA-FA and SIA, and consistent across other pairs) to the baseline model during this FA training stage. It is crucial to note that this increased cost is only incurred during the FA training process. Once the existing detector is enhanced using our FA strategy, the U-Net is not required for inference. The final enhanced model performs inference using its original architecture, thus incurring no additional computational overhead at inference time compared to its pre-FA version, while benefiting from the improved sensitivity to local artifacts.

V. Conclusion

In this paper, we presented IDCNet, a novel deepfake detection framework that addresses the fundamental limitations of existing methods in capturing local manipulation artifacts. Our approach introduces three key innovations: (1) an image decomposition strategy that explicitly separates global content from local details, enabling specialized processing of different forgery cues; (2) a cross-view distillation mechanism that facilitates bidirectional knowledge transfer between complementary views; and (3) a lightweight feature alignment method that enhances existing detection models without architectural modifications. Through extensive experiments, we demonstrated that our approach consistently outperforms state-of-the-art methods across various benchmark datasets while maintaining strong robustness against image perturbations. The success of our method validates our key insight that decomposing images into complementary views enables more effective capture and integration of both global and local forgery features. Furthermore, the practical utility of our pre-trained decomposition model in enhancing existing detectors suggests a promising direction for upgrading deployed detection systems.

References

[1] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., Red Hook, NY, USA: Curran Associates, 2014, pp. 1–9.

[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.

[6] J. Liu, J. Xie, Y. Wang, and Z.-J. Zha, "Adaptive texture and spectrum clue mining for generalizable face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 1922–1934, 2024.

[7] J. Wang, Y. Sun, and J. Tang, "LiSiam: Localization invariance Siamese network for deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2425–2436, 2022.

[8] R. Wang et al., "FacialPulse: An efficient RNN-based depression detection via temporal facial landmarks," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 311–320.

[9] J. Huang et al., "KeystrokeSniffer: An off-the-shelf smartphone can eavesdrop on your privacy from anywhere," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 6840–6855, 2024.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[11] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[12] S. Das, S. Seferbekov, A. Datta, S. Islam, and Md. R. Amin, "Towards solving the DeepFake problem: An analysis on improving deepfake detection using dynamic face augmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3769–3778.

[13] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3994–4004.

[14] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," 2018, *arXiv:1811.00656*.

[15] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5001–5010.

[16] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Germany: Springer, 2020, pp. 86–103.

[17] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16317–16326.

[18] B. Huang et al., "Implicit identity driven deepfake face swapping detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4490–4499.

[19] D. Nguyen et al., "LAA-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17395–17405.

[20] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Munich, Germany: Springer, 2015, pp. 234–241.

[22] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5102–5112.

[23] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.

[24] K. Yang, T. Zhou, Y. zhang, X. Tian, and D. Tao, "Class-disentanglement and applications in adversarial detection and defense," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 16051–16063.

[25] Z. Kuang, C. He, Y. Huang, X. Ding, and H. Li, "Joint image and feature levels disentanglement for generalizable vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15259–15273, Dec. 2023.

[26] K. Shiohara, X. Yang, and T. Taketomi, "BlendFace: Re-designing identity encoders for face-swapping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 7634–7644.

[27] M. Li et al., "Accurate, secure, and efficient semi-constrained navigation over encrypted city maps," *IEEE Trans. Dependable Secure Comput.*, vol. 22, no. 3, pp. 2642–2658, May 2025.

[28] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7184–7193.

[29] M. Li, Y. Chen, C. Lal, M. Conti, F. Martinelli, and M. Alazab, "Nereus: Anonymous and secure ride-hailing service based on private smart contracts," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 4, pp. 2849–2866, Jul./Aug. 2023.

[30] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards high fidelity and occlusion aware face swapping," 2019, *arXiv:1912.13457*.

[31] M. Li, Y. Chen, C. Lal, M. Conti, M. Alazab, and D. Hu, "Eunomia: Anonymous and secure vehicular digital forensics based on blockchain," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 1, pp. 225–241, Jan. 2023.

[32] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[33] P. Zhao, J. Zhou, Y. Zhao, D. Guo, and Y. Chen, "Multimodal class-aware semantic enhancement network for audio-visual video parsing," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 10, pp. 10448–10456.

[34] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019.

[35] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos, "HyperReenact: One-shot reenactment via jointly learning to refine and retarget faces," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 7149–7159.

[36] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[37] M. Liu et al., "STGAN: A unified selective transfer network for arbitrary image attribute editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3673–3682.

[38] J. Chen et al., "PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," 2023, *arXiv:2310.00426*.

[39] S. Atito, M. Awais, and J. Kittler, "SiT: Self-supervised vIsion transformer," 2021, *arXiv:2104.03602*.

[40] J. Gao et al., "StyleShot: A snapshot on any style," 2024, *arXiv:2407.01414*.

[41] J. Gao et al., "FaceShot: Bring any character into life," 2025, *arXiv:2503.00740*.

[42] X. Tian et al., "Variational distillation for multi-view learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 4551–4566, Jul. 2024.

[43] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[44] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[45] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A comprehensive benchmark of deepfake detection," 2023, *arXiv:2307.01426*.

[46] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8692–8701.

[47] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2307–2311.

[48] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5781–5790.

[49] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao, "CORE: Consistent representation learning for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 12–21.

[50] H. Liu et al., "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.

[51] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4113–4122.

[52] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering common features for generalizable deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22412–22423.

[53] Z. Yan et al., "DF40: Toward next-generation deepfake detection," 2024, *arXiv:2406.13495*.

[54] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8060–8069.

[55] R. Xia, D. Liu, J. Li, L. Yuan, N. Wang, and X. Gao, "MMNet: Multi-collaboration and multi-supervision network for sequential deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 3409–3422, 2024.

[56] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 15023–15033.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[58] Z. Guo, Z. Jia, L. Wang, D. Wang, G. Yang, and N. Kasabov, "Constructing new backbone networks via space-frequency interactive convolution for deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 401–413, 2024.

[59] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3207–3216.

[60] B. Dolhansky et al.., "The deepfake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.

[61] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[62] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3677–3685.

[63] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," 2019, *arXiv:1910.12713*.

[64] A. Paszke et al., "An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 1912, p. 8026.

[65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[66] A. Luo, C. Kong, J. Huang, Y. Hu, X. Kang, and A. C. Kot, "Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 1168–1182, 2023.

[67] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[68] P. Jiang, H. Xie, L. Yu, G. Jin, and Y. Zhang, "Exploring bi-level inconsistency via blended images for generalizable face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 6573–6588, 2024.

[69] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2889–2898.

Zhiyuan Wang is currently pursuing the Ph.D. degree with Hefei University of Technology, Hefei, China.

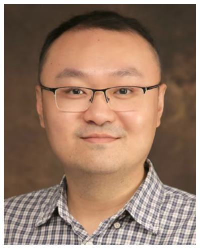
His current research interests include multi-modal information fusion and deepfake detection.

Yanxiang Chen (Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Science and Technology of China in 2004. From 2006 to 2008, she was supported by CSC as a Visiting Scholar collaborating with Prof. Thomas Huang (Foreign Member of Chinese Academy of Sciences) at the University of Illinois at Urbana–Champaign (UIUC), USA, and a Visiting Scholar collaborating with Prof. Shuicheng Yan (IEEE Fellow) at the National University of Singapore (NUS), Singapore, from 2012 to 2013. She is currently a Professor at the School of Computer Science and Information Engineering, Hefei University of Technology. Her research interests include multi-modal signal processing, multimedia content security, pattern recognition, and machine learning. She is a Member of ACM, CCF, CSIG, and CAAI.

Yuanzhi Yao (Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China in 2017. He is currently an Associate Professor with Hefei University of Technology. His research interests include multimedia security and machine learning.

Meng Han (Senior Member, IEEE) received the Ph.D. degree in computer science from Georgia State University, Atlanta, GA, USA, in 2017, and the MBA degree from Georgia Institute of Technology, Atlanta, in 2021. He is currently the Director of the Intelligent Fusion Research Center (IFRC) and a Researcher at Zhejiang University. His research interests include data-driven intelligence, data security and privacy, and AI governance. He is an IEEE COMSOC Member and an ACM Member.

Wenpeng Xing, photograph and biography not available at the time of publication.



Meng Li (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China, in 2019. He was a Post-Doctoral Researcher at the Department of Mathematics and HIT Center, University of Padua, Italy, where he is with the Security and PRIvacy Through Zeal (SPRITZ) Research Group led by Prof. Mauro Conti (IEEE Fellow). He is an Associate Professor and the Personnel Secretary of the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), China. He was sponsored by ERCIM Alain Bensoussan' Fellowship Programme (from October 2020 to March 2021) to conduct post-doctoral research supervised by Prof. Fabio Martinelli at CNR, Italy. He was sponsored by China Scholarship Council (CSC) as a Joint Ph.D. Student (from September 2017 to August 2018) supervised by Prof. Xiaodong Lin (IEEE Fellow) in the Broadband Communications Research (BBCR) Laboratory, University of Waterloo, and Wilfrid Laurier University, Canada. He is supported by CSC as a Visiting Scholar (from March 2025 to June 2025) collaborating with Prof. Mauro Conti at the HIT Center, University of Padua, Italy. His research interests include security, privacy, applied cryptography, blockchain, TEE, and internet of vehicles. In this area, he has published 119 papers in topmost journals and conferences, including IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, TODS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON SERVICES COMPUTING, COMST, IEEE S&P, USENIX Security, ACM MobiCom, and ISSTA.

Dr. Li is a Senior Member of CIE, CIC, and CCF. He has served as a TPC Member for conferences, including ICDCS, Inscrypt, ICICS, and TrustCom. He was a recipient of the 2024 IEEE HITC Award for Excellence (Early Career Researcher) and the 2025 IEEE TCSVC Rising Star Award. He is an Associate Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT.