frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 18 November 2025
DOI 10.3389/fdata.2025.1669488

OPEN ACCESS

EDITED BY
Chen Wang,
Huazhong University of Science and Technology, China

REVIEWED BY
Hadeel Taher,
University of Anbar, Iraq
Xinmiao Ding,
Shandong Technology and Businesses University, China

*CORRESPONDENCE
Xi Wu
xi.wu@cuit.edu.cn
Xin Wang
xwang56@albany.edu

RECEIVED 25 July 2025
ACCEPTED 18 September 2025
PUBLISHED 18 November 2025

CITATION
Yang S, Guo H, Hu S, Zhu B, Fu Y, Lyu S, Wu X and Wang X (2025) CrossDF: improving cross-domain deepfake detection with deep information decomposition.
Front. Big Data 8:1669488.
doi: 10.3389/fdata.2025.1669488

CrossDF: improving cross-domain deepfake detection with deep information decomposition

Shanmin Yang[1], Hui Guo[2], Shu Hu[3], Bin Zhu[4], Ying Fu[1], Siwei Lyu[2], Xi Wu[1]* and Xin Wang[5]*

[1]Computer Science and Technology, Chengdu University of Information Technology, Chengdu, China, [2]University at Buffalo, State University of New York (SUNY), Buffalo, NY, United States, [3]Purdue University, West Lafayette, IN, United States, [4]Microsoft Research Asia, Beijing, China, [5]University at Albany, State University of New York (SUNY), Albany, NY, United States

Deepfake technology represents a serious risk to safety and public confidence. While current detection approaches perform well in identifying manipulations within datasets that utilize identical deepfake methods for both training and validation, they experience notable declines in accuracy when applied to cross-dataset situations, where unfamiliar deepfake techniques are encountered during testing. To tackle this issue, we propose a Deep Information Decomposition (DID) framework to improve Cross-dataset Deepfake Detection (CrossDF). Distinct from most existing deepfake detection approaches, our framework emphasizes high-level semantic attributes instead of focusing on particular visual anomalies. More specifically, it intrinsically decomposes facial representations into deepfake-relevant and unrelated components, leveraging only the deepfake-relevant features for classification between genuine and fabricated images. Furthermore, we introduce an adversarial mutual information minimization strategy that enhances the separability between these two types of information through decorrelation learning. This significantly improves the model's robustness to irrelevant variations and strengthens its generalization capability to previously unseen manipulation techniques. Extensive experiments demonstrate the effectiveness and superiority of our proposed DID framework for cross-dataset deepfake detection. It achieves an AUC of 0.779 in cross-dataset evaluation from FF++ to CDF2 and improves the state-of-the-art AUC significantly from 0.669 to 0.802 on the diffusion-based Text-to-Image dataset.

KEYWORDS

deepfake detection, deep information decomposition, model generalization, decorrelation learning, cross-dataset

1 Introduction

Recent advances in deep generative models, exemplified by Face2Face (Thies et al., 2016), DeepFake (Rossler et al., 2019), and generative adversarial networks (GANs) (Karras et al., 2019), have significantly elevated the visual realism of synthetic facial imagery. While these technologies hold promise in creative and educational domains, their potential for malicious use poses substantial threats to digital security and undermines public trust. A fundamental challenge in current deepfake detection research lies in the pronounced performance degradation encountered in cross-dataset scenarios, where models trained on one forgery technique fail to generalize to others due to domain shift and overfitting to method-specific artifacts.

In response to this challenge, this paper aims to enhance the generalization capability of deepfake detection across diverse manipulation methods. Our primary contribution is a novel framework that explicitly disentangles deepfake-related features from irrelevant technique-specific information and irrelevant identity-related variations, thereby addressing the critical problem of model overfitting to technique-specific artifacts. This approach represents a significant departure from existing methods and offers a more robust solution for real-world deployment.

Although substantial efforts have been made toward detecting deepfakes in recent years and promising performance have been achieved in intra-dataset settings where both training and testing images are generated using the same deepfake technique. Most of the existing approaches rely on identifying specific visual artifacts produced during the deepfake generation process, such as inconsistencies in blending boundaries between genuine and manipulated faces (Li L. et al., 2020), deviations in head poses (Yang et al., 2019), affine face warping distortions (Li and Lyu, 2019), eye-state anomalies (Li et al., 2018), spectral discrepancies (Li et al., 2021), inter-frame illumination inconsistency (Zhu et al., 2024). Consequently, these methods tend to overfit to the unique artifacts of a particular deepfake technique and exhibit limited generalization capability when exposed to unseen manipulation techniques or datasets. For example, the detector proposed by Qian et al. (2020) achieves an AUC score of 0.98 when trained and tested within the same FaceForensics++ (FF++) deepfake dataset (Rossler et al., 2019), but its performance drops significantly to 0.65 (Nadimpalli and Rattani, 2022; Kim and Kim, 2022) when the model is evaluated on the Celeb-DF dataset (Li Y. et al., 2020) under cross-dataset protocols.

In analyzing the issue of cross-dataset performance degradation, we observe that deepfake detection constitutes a form of fine-grained image classification. As deepfake generation techniques continue to evolve, the discrepancies between authentic and manipulated images have become increasingly subtle, often more nuanced than the variations among deepfakes produced from the same source image using different forgery methods. Furthermore, features directly extracted with conventional deep neural networks (e.g., EfficientNet Tan and Le, 2021) from deepfake images often encapsulate entangled representations that intertwine forgery-related artifacts, domain-specific attributes of the manipulation method, and identity-related factors such as facial expressions and appearance. As illustrated in Figure 1, this feature entanglement exacerbates the sensitivity of detection models to variations in irrelevant factors, particularly those that dominate the representation, thereby impairing generalization across domains.

Inspired by these insights, we propose a Deep Information Decomposition (DID) framework for cross-dataset deepfake detection, as illustrated in Figure 2. Unlike traditional methods that rely on low-level visual artifacts, our approach emphasizes high-level semantic features to capture more generalized forgery cues. Specifically, we regard face images generated by different deepfake techniques, such as Face2Face (Thies et al., 2016) and DeepFake (Rossler et al., 2019), as distinct data domains, thereby formulating cross-dataset detection as a domain generalization problem.

Within the proposed framework, facial representations are adaptively decomposed into three semantically distinct components: deepfake-related information, which captures universal manipulation traces common across different forgery

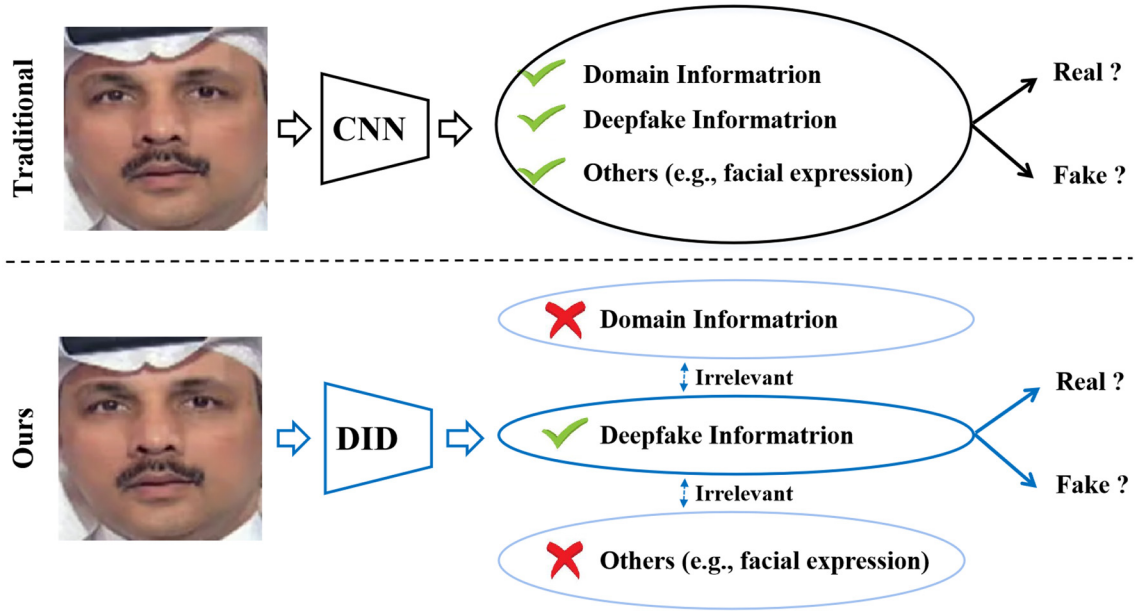

FIGURE 1
Various information changes entangled with the deepfake information in traditional methods **(top)** would affect real/fake classification accuracy, leading to a sharp degradation in performance when the discrepancies in these components between the training and test sets are more significant than the differences between real and deepfake information. Our deep information decomposition (DID) method **(bottom)** separates the deepfake information from various information irrelevant to real/fake classification to improve the robustness of deepfake detection (FF++ dataset image adapted with permission from Rossler, A., 2019, https://www.kaggle.com/datasets/xdxd003/ff-c23, licensed under CC-BY-NC).

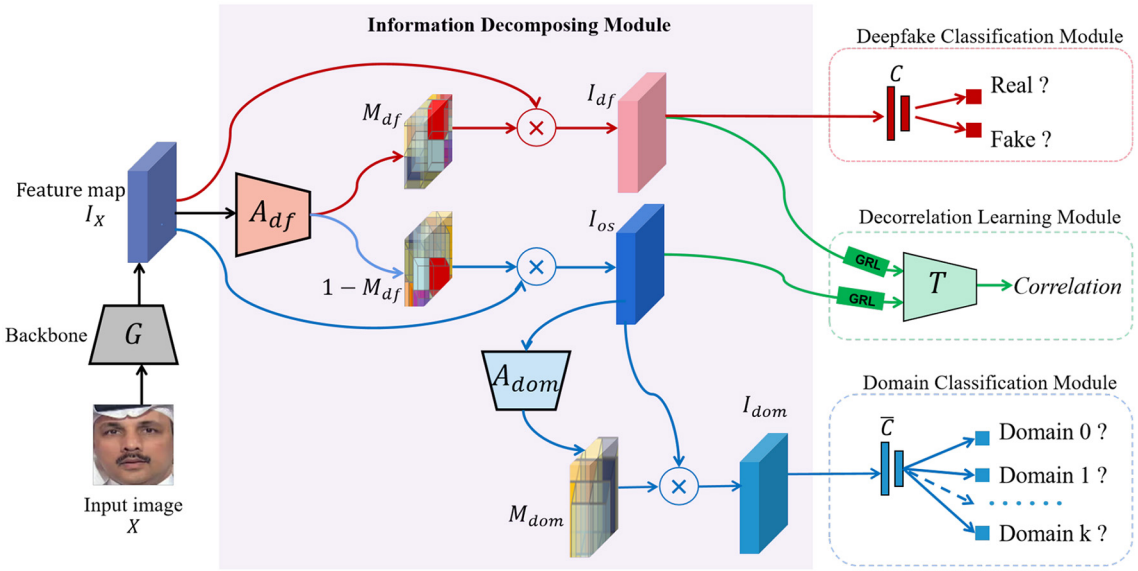

FIGURE 2
Overview of our Deep Information Decomposition (DID) framework: the feature map IX of an input facial image X, generated by a backbone network G, is adaptively split into deepfake information I_{df} and non-deepfake information I_{os}, under the guidance of the deepfake attention network A_{df} and the supervision of the deepfake classification module (FF++ dataset image adapted with permission from Rossler, A., 2019, https://www.kaggle.com/datasets/xdxd003/ff-c23, licensed under CC-BY-NC). The domain attention network A_{dom} and the domain classification module extract the forgery method-related (domain) information I_{dom}, ensuring that I_{dom} is included in the non-deepfake information I_{os} but being excluded from the deepfake information. Additionally, the decorrelation learning module enforces no overlap between deepfake information and non-deepfake information. This module consists of an information estimation network T, which functions in a max-min manner with the information decomposition module through the gradient reversal layer (GRL). C and C are the deepfake and domain classifiers, respectively.

methods; irrelevant technique-specific information, which encodes distinctive artifacts attributable to particular generation approaches such as Face2Face (Thies et al., 2016) and DeepFake (Rossler et al., 2019); and irrelevant identity-related variations, including facial expressions, lighting conditions, and personal identity attributes. This decomposition is achieved through two dedicated attention modules that learn to selectively emphasize and isolate these components in a self-supervised manner. Crucially, only the deepfake-relevant features are utilized during the authentic-versus-fake classification phase, effectively suppressing the influence of extraneous factors and technique-specific biases. To further enhance feature discriminability and domain invariance, we introduce a decorrelation learning module that minimizes mutual information between the deepfake-specific features and both types of irrelevant variations (identity-related and technique-specific) via adversarial training. This explicitly encourages statistical independence among feature components, thereby significantly improving the model's robustness and generalization capability across unseen datasets and diverse manipulation techniques. Extensive experiments on multiple benchmarks demonstrate the effectiveness and superiority of our framework in cross-dataset deepfake detection scenarios.

In summary, our main contributions are:

1. We propose a novel end-to-end Deep Information Decomposition (DID) framework. It formulates cross-dataset deepfake detection as a domain generalization problem and decomposes face image information into deepfake-related, technique-specific, and identity-related information components to enhance generalization.

2. We introduce a decorrelation learning module that promotes the independence of the deepfake-related information from all irrelevant variations (identity-related and technique-specific) without requiring knowledge of their distribution functions or relationships, thereby enhancing the robustness of deepfake detection.

3. We conducted extensive experiments that demonstrated the superiority of our framework, achieving state-of-the-art performance on the challenging cross-dataset deepfake detection task.

2 Related work

This section provides a brief review of deepfakes, cross-dataset deepfake detection, and information decomposing. For more details about the deepfake techniques and deepfake datasets, please refer to Nguyen et al. (2022).

2.1 Deepfakes

Deepfake, broadly referring to manipulated or synthetic media that convincingly mimics natural content (Nguyen et al., 2022), poses a significant threat to digital media integrity. This paper focuses specifically on deepfake faces. Existing creation methods can be broadly classified into two categories: transfer-based and synthesis-based approaches.

Transfer-based deepfake methods manipulate target faces by transferring facial attributes (e.g., expression, mouth movement)

or entire identities from a reference source. The primary benefit of these methods is their ability to produce highly convincing forgeries by leveraging real human features, making the manipulations contextually consistent and realistic. For instance, Face2Face (Thies et al., 2016) enables real-time facial reenactment by transferring expressions from a source to a target video, preserving the target's identity. FaceSwap and DeepFake (Rossler et al., 2019) replace the face region in a target video with a source face, often relying on autoencoders to learn and swap identity attributes. Neural Textures (Thies et al., 2019) combines learned neural textures with deferred rendering to improve the realism of synthesized facial motions, particularly around critical areas like the mouth. Similarly, Bao et al. (2018) decomposes faces into identity and attribute representations, allowing controlled attribute transfer while retaining original identity. However, a key limitation of these methods is their reliance on blending operations and warping, which often introduce subtle artifacts. These can manifest as inconsistencies in lighting, blurring at blending boundaries, misaligned facial geometry, or unnatural eye movements. While these artifacts form the basis for many early detection algorithms, they also represent a point of vulnerability for forgers: as generation models improve, these artifacts become increasingly subtle, making detection more challenging.

Synthesis-based deepfake methods, in contrast, generate entirely novel facial images or attributes without direct reference to a specific source individual. This category is dominated by Generative Adversarial Networks (GANs) and variants of 3D Morphable Models (3DMM). The main advantage of these approaches is their ability to create highly diverse and novel forgeries that do not rely on the availability of a reference face/video, expanding the scope of potential attacks. For example, 3DMM-guided approaches (Geng et al., 2019), which generate arbitrary facial expressions and viewpoints under structured geometric control, enhancing pose robustness. StyleGAN (Karras et al., 2019), which produces high-fidelity, diverse facial imagery through its style-based generator, significantly raising the visual quality bar for synthetic faces. GANprintR (Neves et al., 2020), which is specifically designed to generate realistic deepfakes while attempting to evade detection by minimizing known GAN fingerprints. Despite their high visual quality, synthesis-based methods can introduce their own unique artifacts. These include frequency domain abnormalities (e.g., spectral disparities), physiological implausibilities (e.g., asymmetric pupils), and inherent fingerprints left by the generator architecture itself. Nevertheless, the rapid advancement of these technologies has led to a continuous reduction of such artifacts, rendering detection strategies that rely on them increasingly obsolete.

2.2 Cross-dataset deepfake detection

The paramount challenge in contemporary deepfake detection is generalization, the ability of a model to perform robustly on forgeries generated by unseen methods or datasets. While numerous detection methods (Zhao H. et al., 2021; Shiohara and Yamasaki, 2022; Dong et al., 2022) achieve impressive performance in intra-dataset scenarios, they suffer from catastrophic performance degradation under cross-dataset evaluation. This failure mode primarily stems from models overfitting to technique-specific artifacts (e.g., blending patterns of FaceSwap, frequency signatures of a specific GAN) rather than learning a universal representation of "forgery". In response to this generalization challenge, research has evolved along several key directions: Data Augmentation for Domain Expansion, a straightforward strategy is to augment training data to simulate a wider variety of forgery types, thereby encouraging the model to learn more invariant features. For instance, Zhao T. et al. (2021) proposed dynamic data augmentation strategies to artificially expand the diversity of training samples, effectively exposing the model to a broader spectrum of potential artifacts. Nadimpalli and Rattani (2022) advanced this concept by employing a reinforcement learning-based strategy to intelligently select augmentation policies, mitigating domain shift more effectively than random strategies. A core issue with these approaches is that they rely on synthetic augmentations which may not fully capture the complex and realistic distribution of novel deepfake techniques, potentially limiting their effectiveness against truly advanced unseen forgeries.

Inherent Forensic Feature Learning, another line of work seeks to identify and leverage common, intrinsic traces left by deepfake generation processes that are theoretically invariant across different methods. Kim and Kim (2022) focused on color distribution inconsistencies introduced during the face-synthesis process, a low-level cue that is often shared across different manipulation techniques. Yu et al. (2022) aimed to learn and amplify common forgery features that persist across diverse datasets, moving beyond dataset-specific biases. The fundamental challenge here is that as generative models become more advanced, they produce fewer inherent artifacts. This makes the discovery of robust, shared forensic features increasingly difficult.

Explicit Domain Alignment and Bridging, the most directly relevant approaches explicitly model and aim to reduce the distributional gap between different deepfake domains. Yu et al. (2023) made significant strides by using Adaptive Normalization layers and generating "bridging samples" to create a continuous latent space between domains, explicitly narrowing the distribution gap for improved generalization. Similarly, Yin et al. (2024) employed a framework based on Invariant Risk Minimization (IRM), designed to prioritize domain-invariant features and aligned representations, thus enhancing cross-domain performance. Huang et al. (2023) introduced a video-level contrastive learning framework to maintain feature consistency across varying compression levels, a critical step toward real-world applicability where compression is ubiquitous. While effective, many of these methods can introduce significant algorithmic complexity (e.g., requiring multiple domains during training, generative modules for sample synthesis, or complex loss functions), which may hinder their practical deployment and scalability.

While existing research has made valuable progress through data augmentation, feature learning, and domain alignment, the problem remains largely open. Many methods still struggle with the sheer diversity and evolving nature of deepfake techniques, often requiring complex multi-domain training or failing to generalize to the next generation of generators. This underscores the need for a

more elegant and principled approach to learning domain-agnostic forgery features. Our proposed method addresses this by explicitly disentangling deepfake-related features from irrelevant variations, aiming to isolate a more pure and generalizable representation of manipulation that is invariant to the creation method.

2.3 Information decomposing

Information decomposition, which aims to disentangle complex and intertwined data into distinct, semantically meaningful components and isolate those relevant to specific tasks, has been widely adopted across various computer vision applications. For instance, methods such as those proposed by Tran et al. (2017) and Wang et al. (2019) separate identity-related features from pose and age variations, respectively, thereby reducing the influence of these factors in face recognition systems. Similarly, Wu et al. (2019) decomposes facial representations into identity and modality components to improve performance in NIR-VIS heterogeneous face recognition.

In the domain of deepfake detection, several studies have leveraged disentanglement strategies to enhance generalization and detection accuracy. Hu et al. (2021) detect forgery regions by disentangling multi-scale features and training the detector specifically on these localized artifacts. Liang et al. (2022) separate artifact-related features from content information to minimize the confounding effect of identity and background during detection. More recently, Yu et al. (2024) introduced a framework that progressively disentangles forgery-relevant features from source-related features through multi-view learning, operating from image space to feature space. Likewise, Yan et al. (2023) employed a multi-task learning setup with a conditional decoder to isolate generalizable forgery attributes from those specific to particular generation methods.

In this paper, we propose an information decomposition framework that achieves disentanglement using a complementary attention mechanism, differing from methods such as Hu et al. (2021); Liang et al. (2022); Yan et al. (2023), which achieve information disentanglement through feature encoding and decoding with carefully designed reconstruction losses (e.g., self-reconstruction, cross-reconstruction, and feature reconstruction). Furthermore, we introduce a deep decorrelation module to ensure that the forgery-relevant features used for deepfake detection remain inherently independent of other features. This intrinsic independence is a crucial factor that is frequently neglected in existing literature like Yu et al. (2024), where such independence is not emphasized. Our strategy for achieving feature independence diverges from that of Yan et al. (2023), who employ a contrastive loss to optimize the Euclidean distance between the decoupled features. Our approach enhances the model's robustness and generalization across different forgery techniques and datasets.

3 Our method

The pipeline of our proposed method is illustrated in Figure 2. Specifically, for an input image X, we employ a CNN-based feature extractor G parameterized by θ to capture its representative features, denoted as $I_X := G(\theta; X)$. These extracted features are then decomposed into three components: (1) the deepfake-related representation, I_{df}, which contains the critical information used to detect deepfakes; (2) the domain-related representation, I_{dom}, which captures the characteristics of the forgery technique or method responsible for generating the deepfake; and (3) the remaining representation. The information decorrelation module ensures that the deepfake information I_{df} is optimized to be independent of other representations, thereby enhancing the performance of the decomposition. The robust deepfake classification module is designed to train a model capable of classifying deepfakes effectively, even in imbalanced datasets, thus enhancing the model's generalization ability. Additionally, the domain classification module is intended to identify the domain to which I_{dom} belongs. Before diving into the details of these modules, we introduce some commonly used notations.

3.1 Notation

Our method takes images from existing deepfake datasets as input data. Let $\mathcal{S} = \{(X_i, Y_i, D_i)\}_{i=1}^n$ be a training dataset that contains images $X_i \in \mathbb{R}^d$ and their corresponding labels $Y_i \in \{0, 1\}$, where 0 denotes real and 1 indicates fake. $D_i := [D_i^0, D_i^1, ..., D_i^k]^\top$ represents the domain label of X_i, where the domain size of fake data $k \geq 1$ and $D_i^j \in \{0, 1\}, \forall j \in \{0, 1, ..., k\}$. In particular, $D_i^j = 1$ indicates that X_i is from the j-th domain. Specifically, X_i is from the real data domain if $j = 0$ and from the fake data domain j (i.e., forged by the method j) if $j > 0$. For example, the fake images in the FF++ dataset (Rossler et al., 2019) are generated by four face manipulation methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Therefore, $k = 4$. In this work, we assume each image comes from a single domain.

3.2 Information decomposition module

Motivated by Yang et al. (2021), the information decomposition module consists of a deepfake attention network A_{df} parameterized by ψ [denoted as $A_{df}(\psi; .)$] and a domain attention network A_{dom} parameterized by φ [denoted as $A_{dom}(\varphi; .)$], as shown in Figure 2. Taking the face information I_X embedded with entangled information as input, the deepfake attention network focuses on deepfake-relevant information, thereby it decomposes I_X into two complementary components: the deepfake-relevant information I_{df} and the deepfake-irrelevant information I_{os}. This process can be formulated as follows,

$$\begin{aligned} M_{df} &= A_{df}(\psi; I_X), \\ I_{df} &= M_{df} \otimes I_X, \\ I_{os} &= (1 - M_{df}) \otimes I_X, \end{aligned} \tag{1}$$

where $M_{df} \in [0, 1]^{c \times h \times w}$ is the deepfake-relevant information attention map; $\otimes$ represents the Hadamard product.

After receiving the deepfake-irrelevant information I_{os}, the domain attention network A_{dom} focuses on extracting and modeling explicitly forgery technique information. It decomposes

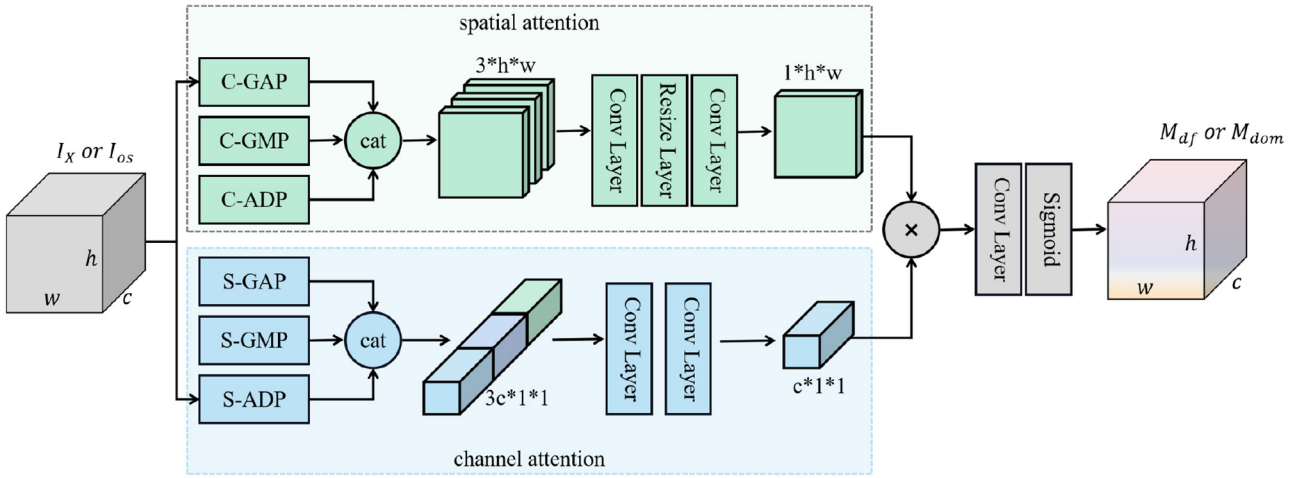

FIGURE 3
Architecture of the deepfake (domain) attention network. This network takes the face information I_X (deepfake-irrelevant information I_{os}) as input and then learns to produce an attention map that highlights the significance (potential) of the input data being correlated with deepfake-relevant (domain-relevant) information. "cat" means concatenating all input data along the channel dimension; ⊗ represents the Hadamard product; "C-GAP," "C-GMP," and "C-ADP" represent cross-channel global average pooling, cross-channel global max pooling, and cross-channel adaptive pooling, respectively; "S-GAP," "S-GMP," and "S-ADP" are cross-spatial global average pooling, cross-spatial global max pooling, and cross-spatial adaptive pooling, respectively.

the deepfake-irrelevant information I_{os} into the forgery technique-related information I_{dom} and others as follows,

$$M_{dom} = A_{dom}(\varphi; I_{os}),$$
$$I_{dom} = M_{dom} \otimes I_{os}, \tag{2}$$

where $M_{dom} \in [0,1]^{c\times h\times w}$ is the forgery technique-related information attention map.

The deepfake attention network A_{df} and the domain attention network A_{dom} are constructed to learn attention maps across spatial and channel dimensions concurrently. They share an identical architecture as shown in Figure 3, where each convolution (Conv) Layer is followed by a PReLU (Parametric Rectified Linear Unit, PReLU) activation function. S-ADP is implemented by a channel-wise spatial convolution layer followed by a sum pooling layer, while C-ADP is implemented with a 1×1 convolution layer. Both A_{df} and A_{dom} are trained to efficiently capture the essential deepfake-related and domain-related information within the input data, respectively. The pseudocode of information decomposition is shown in Algorithm 1.

Input: Entangled face information I_X, deepfake attention network $A_{df}(\psi; .)$, and domain attention network $A_{dom}(\varphi; .)$
Output: Decomposed components I_{df}, I_{dom}, and I_{res}
Step 1: Decompose deepfake-relevant information
$M_{df} \leftarrow A_{df}(\psi; I_X)$
$I_{df} \leftarrow M_{df} \otimes I_X$
$I_{os} \leftarrow (1 - M_{df}) \otimes I_X$
Step 2: Decompose technique-specific information
$M_{dom} \leftarrow A_{dom}(\varphi; I_{os})$
$I_{dom} \leftarrow M_{dom} \otimes I_{os}$
$I_{res} \leftarrow (1 - M_{dom}) \otimes I_{os}$

Algorithm 1. Information decomposition process.

3.3 Decorrelation learning module

The disentangled elements (deepfake and non-deepfake information) are anticipated to be partitioned into two separate representations. To accomplish this, orthogonal constraints are commonly applied to these disentangled components (Wang et al., 2019). However, linear dependence/independence can hardly characterize the intricate relationships between deepfake and non-deepfake information in a high-dimensional and non-linear space. In contrast, mutual information (Kinney and Atwal, 2014) (MI) is capable of capturing arbitrary dependencies between any two variables.

With this motivation, we apply mutual information to evaluate dependencies between deepfake information I_{df} and non-deepfake information I_{os}, formulated as follows:

$$\mathrm{MI}(I_{df}; I_{os}) = \mathbb{D}_{\mathrm{KL}}(P(I_{df}, I_{os}) || P(I_{df}) \otimes P(I_{os})), \tag{3}$$

where $P(\cdot,\cdot)$ is the joint probability distribution, $P(\cdot)$ denotes the marginal probability distribution, and $\mathbb{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence (Joyce, 2011).

Since the probability densities $P(I_{df}, I_{os})$ and $P(I_{df}) \otimes P(I_{os})$ are not known, it becomes challenging to directly minimize $\mathrm{MI}(I_{df}; I_{os})$. Belghazi et al. (2018) introduced a Mutual Information Neural Estimation (MINE) to derive a lower bound on MI's Donsker-Varadhan representation. Subsequently, Hjelm et al. (2019) proposed a Jensen-Shannon MI estimator, which is based on the Jensen-Shannon divergence (Menéndez et al., 1997). This

method has been demonstrated to be more stable and yields improved results.

Inspired by Hjelm et al. (2019), we construct a mutual information estimation network T with parameterizes ϕ to approximate $\mathrm{MI}(I_{df}; I_{os})$ as follows,

$$\begin{aligned}\mathrm{MI}(I_{df}; I_{os}) &\geq \hat{I}^{JSD}(I_{df}; I_{os}) \\ &= \mathbb{E}_{x\sim P(I_{df},I_{os})}\left[\log\sigma(T(\phi; x))\right] \\ &+ \mathbb{E}_{x\sim P(I_{df})\otimes P(I_{os})}\left[\log\left(1-\sigma(T(\phi; x))\right)\right],\end{aligned} \tag{4}$$

where σ is the sigmoid function; $T(\phi;\cdot):\mathbb{R}^{d_x} \rightarrow \mathbb{R}$ acts as the discriminator function in GANs (d_x is the dimension of I_{df} and I_{os}), it aims to estimate and maximize the lower bound of $\mathrm{MI}(I_{df}; I_{os})$, while the target of the previously designed information decomposition module (acting as the generator function in GANs) is to minimize the MI value between I_{df} and I_{os} to achieve a sufficient separation. Specifically, we have the following learning objectives:

$$\begin{aligned}\mathcal{L}_{dec} = \min_{\theta,\psi}\max_{\phi}(&\mathbb{E}_{x\sim P(I_{df},I_{os})}\left[\log\sigma(T(\phi; x))\right] \\ &+\mathbb{E}_{x\sim P(I_{df})\otimes P(I_{os})}\left[\log\left(1-\sigma(T(\phi; x))\right)\right]).\end{aligned} \tag{5}$$

To implement the aforementioned min-max game using standard back-propagation (BP) training, we incorporate a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015) before the network T (as illustrated in Figure 2). During the back-propagation procedure, the GRL modifies the gradient by multiplying a negative scalar, $-\beta$, as it passes from the subsequent layer to the preceding layer, where we typically set $\beta \in (0, 1)$. This technique has also been utilized in various existing works, such as Belghazi et al. (2018); Hjelm et al. (2019). The network T consists of three convolution layers, each followed by a ReLU activation function, and concludes with a fully connected (FC) layer.

3.4 Deepfake classification module

After extracting deepfake-related information I_{df}, we need to consider how to leverage it for training a deepfake detection model. In existing literature, Binary Cross-entropy (BCE) loss is commonly employed for this purpose. However, it is well-known that the BCE loss lacks robustness against imbalanced datasets. Training models with BCE loss on one specific deepfake dataset can lead to a considerable performance decline when evaluated on a different deepfake dataset (Pu et al., 2022).

Building on this observation, we propose a robust deepfake detection loss aimed at enhancing the generalization ability of the trained model by utilizing deepfake-related information I_{df} instead of the complete information I_X. The AUC metric inspires our loss since it is a robust measure to evaluate the classification capability of a model, especially when facing imbalanced data. Specifically, it estimates the size of the area under the receiver operating characteristic (ROC) curve (AUC; He and Garcia, 2009), which is composed of False Positive Rates (FPRs) and True Positive Rates (TPRs). However, the AUC metric cannot be directly used as a loss function since it is challenging to compute during each training iteration. Inspired by Pu et al. (2022), we use the normalized WMW statistic (Yan et al., 2003), equivalent to AUC, to design our loss function.

Specifically, we define a set of indices of fake instances and real instances as $\mathcal{F} = \{i|Y_i = 1\}$ and $\mathcal{R} = \{i|Y_i = 0\}$, respectively. We add a multilayer perceptron (MLP) $C:\mathbb{R}^{d_x} \rightarrow \mathbb{R}$ (d_x is the dimension of I_{df}) parameterized by ω to distinguish fake and real instances, where the input is I_{df} and the output is a real value. Network C predicts input I_{df} to be fake with probability $\sigma(C(\omega; I_{df}))$. Without loss of generality, $C(\omega; I_{df})$ induces the prediction rule such that the predicted label of I_{df} can be $\mathbb{I}[\sigma(C(\omega; I_{df})) \geq 0.5]$, where $\mathbb{I}[\cdot]$ is an indicator function with $\mathbb{I}[a] = 1$ if a is true and 0 otherwise. For simplicity, we assume $C(\omega; I_{df}^{X_i}) \neq C(\omega; I_{df}^{X_j})$ for any $X_i \neq X_j$ (ties can be broken in any consistent way), where $I_{df}^{X_i}$ represents the deepfake information of the sample X_i. Then the normalized WMW can be formulated as follows,

$$\mathrm{WMW} = \frac{1}{|\mathcal{F}||\mathcal{R}|}\sum_{i\in\mathcal{F}}\sum_{j\in\mathcal{R}}\mathbb{I}[C(\omega; I_{df}^{X_i}) > C(\omega; I_{df}^{X_j})], \tag{6}$$

where $|\mathcal{F}|$ and $|\mathcal{R}|$ are the cardinality of $\mathcal{F}$ and $\mathcal{R}$, respectively. However, WMW is non-differentiable due to the indicator function, which is the primary obstacle to using it as a loss function. Therefore, we use its alternative version (Yan et al., 2003):

$$\mathcal{L}_{\mathrm{AUC}} = \frac{1}{|\mathcal{F}||\mathcal{R}|}\sum_{i\in\mathcal{F}}\sum_{j\in\mathcal{R}}E(C(\omega; I_{df}^{X_i}), C(\omega; I_{df}^{X_j})), \tag{7}$$

with

$$\begin{aligned}&E(C(\omega; I_{df}^{X_i}), C(\omega; I_{df}^{X_j})) \\ &:= \begin{cases}(-(C(\omega; I_{df}^{X_i}) - C(\omega; I_{df}^{X_j}) - \gamma))^p, & C(\omega; I_{df}^{X_i}) - C(\omega; I_{df}^{X_j}) < \gamma, \\ 0, & \text{otherwise},\end{cases}\end{aligned} \tag{8}$$

where $0 < \gamma \leq 1$ and $p > 1$ are two hyperparameters. We combine this AUC loss and the conventional BCE loss to construct a learning objective for robust deepfake classification:

$$\begin{aligned}\mathcal{L}_{cls} &= \alpha\mathcal{L}_{\mathrm{BCE}} + (1-\alpha)\mathcal{L}_{\mathrm{AUC}} \\ \mathcal{L}_{\mathrm{BCE}} &= -\frac{1}{n}\sum_{i=1}^{n}[Y_i \cdot \log(\sigma(C(\omega; I_{df}^{X_i}))) \\ &\quad + (1-Y_i)\cdot\log(1-\sigma(C(\omega; I_{df}^{X_i})))]\end{aligned} \tag{9}$$

where α is a hyperparameter designed to balance the weights of the BCE loss and the AUC loss.

3.5 Domain classification module

A domain classification module is also designed using another MLP $\overline{C}:\mathbb{R}^{d_{I_X}} \rightarrow \mathbb{R}^{k+1}$ parameterized by $\overline{\omega}$ to map the forgery method related-domain information I_{dom} into a (k+1)-dimensional domain vector. Specifically, we have $\overline{C}(\overline{\omega}; I_{dom}) = [\overline{C}^0(\overline{\omega}; I_{dom}), \overline{C}^1(\overline{\omega}; I_{dom}), ..., \overline{C}^k(\overline{\omega}; I_{dom})]^\top$, where $\overline{C}^j(\overline{\omega}; I_{dom})$ is the j-th domain prediction. We then apply the softmax function

to compute the probability of each domain that I_{dom} belongs to and combine its domain label to construct a domain classification loss based on the cross-entropy (CE) loss. Therefore, we have

$$\mathcal{L}_{dom} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k} D_i^j \log(S[\overline{C}^j(\overline{\omega}; I_{dom}^{X_i})]) \tag{10}$$

where $S[\overline{C}^j(\overline{\omega}; I_{dom}^{X_i})] \in (0,1)$ is the j-th domain predicted probability for the domain information of I_{X_i} after using softmax operator $S[\cdot]$.

3.5.1 Overall loss

To sum up, the proposed framework is optimized with the following final loss function:

$$\mathcal{L} = \lambda_{dec}\mathcal{L}_{dec} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{dom}\mathcal{L}_{dom} \tag{11}$$

where λ_{dec}, λ_{cls}, and λ_{dom} are hyperparameters that can balance these loss terms. In practice, the optimization problem in Equation 11 can be solved with an iterative stochastic gradient descent and ascent approach (Beznosikov et al., 2022). Specifically, we first initialize the model parameters θ, ψ, φ, ϕ, ω, and $\overline{\omega}$. Then we alternate uniformly at random a mini-batch $\mathcal{S}_b$ of training samples from the training set $\mathcal{S}$ and do the following steps on $\mathcal{S}_b$ for each iteration:

$$\begin{pmatrix} \theta_{l+1} \\ \psi_{l+1} \\ \varphi_{l+1} \\ \phi_{l+1} \\ \omega_{l+1} \\ \overline{\omega}_{l+1} \end{pmatrix} = \begin{pmatrix} \theta_l \\ \psi_l \\ \varphi_l \\ \phi_l \\ \omega_l \\ \overline{\omega}_l \end{pmatrix} - \begin{pmatrix} \eta_\theta \partial_\theta \mathcal{L}|_{\theta=\theta_l} \\ \eta_\psi \partial_\psi \mathcal{L}|_{\psi=\psi_l} \\ \eta_\varphi \partial_\varphi \mathcal{L}|_{\varphi=\varphi_l} \\ -\eta_\phi \beta \partial_\phi \mathcal{L}|_{\phi=\phi_l} \\ \eta_\omega \partial_\omega \mathcal{L}|_{\omega=\omega_l} \\ \eta_{\overline{\omega}} \partial_{\overline{\omega}} \mathcal{L}|_{\overline{\omega}=\overline{\omega}_l} \end{pmatrix}, \tag{12}$$

where $\mathcal{L}$ is defined on $\mathcal{S}_b$, η_θ, η_ψ, η_φ, η_ϕ, η_ω, and $\eta_{\overline{\omega}}$ are learning rates, and $\partial_\theta \mathcal{L}$, $\partial_\psi \mathcal{L}$, $\partial_\varphi \mathcal{L}$, $\partial_\phi \mathcal{L}$, $\partial_\omega \mathcal{L}$, and $\partial_{\overline{\omega}} \mathcal{L}$ are the (sub)gradient of $\mathcal{L}$ with respect to θ, ψ, φ, ϕ, ω, and $\overline{\omega}$. In the testing phase, we only use the feature extractor G, attention module A_{df}, and the deepfake classification module C. The pseudocode is shown in Algorithm 2.

Input: A training dataset $\mathcal{S}$ of size n, β, max_iterations, num_batch, η_θ, η_ψ, η_φ, η_ϕ, η_ω, and $\eta_{\overline{\omega}}$
Output: A robust Deepfake detector with parameters θ^*, ψ^*, and ω^*

Initialization: θ_0, ψ_0, φ_0, ϕ_0, ω_0, $\overline{\omega}_0$, $l = 0$ **for** $e = 1$ *to* *max_iterations* **do**
  for $b = 1$ *to* *num_batch* **do**
    Sample a mini-batch $\mathcal{S}_b$ from $\mathcal{S}$
    Update parameters with Equation 12.
    $l \leftarrow l + 1$
  end
end
return $\theta^* \leftarrow \theta_l$, $\psi^* \leftarrow \psi_l$, $\omega^* \leftarrow \omega_l$

Algorithm 2. Deep information decomposition.

4 Experiments

This section evaluates the effectiveness of the proposed DID framework in terms of cross-dataset deepfake detection performance. In the following discussion, we will exchange the "method" or "framework" used for DID.

4.1 Experimental settings

4.1.1 Datasets

For fair comparisons with state-of-the-art methods, we adopt the two most widely used datasets, FF++ (Rossler et al., 2019) and Celeb-DF (Li Y. et al., 2020), in our experiments. Specifically, the high-quality (HQ, compressed with a constant rate factor of 23) version of FF++ is utilized throughout all experiments. It includes one real video subset and four fake video subsets generated using FaceSwap, DeepFakes, Face2Face, and Neural Textures techniques, respectively. Each subset contains 1,000 videos, split into 720 for training, 140 for validation, and 140 for testing (Rossler et al., 2019). The Celeb-DF (Li Y. et al., 2020) dataset contains real and fake videos of 59 celebrities. Following the official protocols in Li Y. et al. (2020), we use the latest version, Celeb-DF V2 (CDF2), which includes 590 real celebrity (Celeb-real) videos, 300 real videos from YouTube (YouTube-real) and 5,639 synthesized celebrity (Celeb-synthesis) videos generated from Celeb-real.

4.1.2 Compared methods and evaluation metrics

To assess the performance of our framework, we benchmark it against the following state-of-the-art (SOTA) frame-level baseline methods: F^3-Net (Qian et al., 2020), CFFs (Yu et al., 2022), RL (Nadimpalli and Rattani, 2022), Multi-task (Nguyen et al., 2019), Two Branch (Masi et al., 2020), MDD (Zhao H. et al., 2021), and NoiseDF (Wang and Chow, 2023). The results of Multi-task and Two Branch are drawn from Nadimpalli and Rattani (2022), while these for MDD are referenced from Yu et al. (2022). We evaluate the methods using two commonly used metrics: the area under the receiver operating characteristic curve (AUC) and the equal error rate (EER), both of which are standard in previous works for performance comparison.

4.1.3 Implementation details

In our experiments, we utilize EfficientNet v2-L (Tan and Le, 2021) pre-trained on the ImageNet dataset as the backbone for feature extraction. All face images are aligned to a size of 224 × 224 using the MTCNN method (Zhang et al., 2016), and subsequently converted from RGB to grayscale before being fed into the proposed framework. The model is trained using the Adam optimizer with a weight decay of $5e^{-4}$ and a learning rate of $1e^{-5}$. We set the learning rate η_ψ, η_φ, η_ϕ, η_ω, and $\eta_{\overline{\omega}}$ in Equation 12 to be 10 times that of η_θ ($\eta_\theta = 1e^{-5}$). The batch size is set to 15, and each epoch consists of 6000 iterations. We set γ and p in Equation 8

TABLE 1 Intra-dataset evaluation on FF++ and cross-dataset evaluation from FF++ to CDF2.

Methods	Intra-dataset	Cross-dataset
	AUC ↑	AUC ↑
Xception (Rossler et al., 2019)	0.997	0.653
Multi-task (Nguyen et al., 2019)	0.763	0.543
Two Branch (Masi et al., 2020)	0.931	0.734
MDD (Zhao H. et al., 2021)	**0.998**	0.674
RL (Nadimpalli and Rattani, 2022)	0.994	0.669
F^3-Net (Qian et al., 2020)	0.981	0.651
CFFs (Yu et al., 2022)	0.976	0.742
NoiseDF (Wang and Chow, 2023)	0.940	0.759
FDML (Yu et al., 2024)	0.996	0.731
DIRE (Wang et al., 2023)	0.994	-
DID (Ours)	0.970	**0.779**

The highest results are highlight in bold.

TABLE 2 Cross-dataset evaluation from FF++ to the diffusion-generated DiFF dataset.

Methods	Test subset			
	T2I	I2I	FS	FE
Xception (Rossler et al., 2019)	0.624	0.568	**0.860**	0.586
F^3-Net (Qian et al., 2020)	0.669	0.676	0.810	0.606
DIRE (Wang et al., 2023)	0.442	0.646	0.850	0.577
DID (Ours)	**0.802**	**0.741**	0.817	**0.676**

The highest results are highlight in bold.

to 0.15 and 2, respectively. We use $\alpha = 0.5$ in Equation 9. The hyperparameters λ_{cls}, λ_{dom}, and λ_{dec} in Equation 11 are set to 1, 1, and 0.01, respectively. The hyperparameter β in Equation 12 is adapted to increase from 0 to 1 in the training procedure as $\beta = 2.0/(1.0 + e^{-5p}) - 1.0$, where p is the ratio of the current training epochs to the maximum number of training epochs. All experiments are conducted on two NVIDIA RTX 3080 GPUs, using Pytorch 1.10 and Python 3.6. In all our experiments, no data augmentation techniques, such as random image compression, image flip, and brightness contrast, are employed.

4.2 Intra-dataset evaluation

We assess the detection performance of our proposed method, DID, in the intra-dataset scenario, where both the training and testing datasets are derived from the FF++ dataset and are disjoint from one another. Table 1 presents the results of the intra-dataset evaluation along with comparisons to the baseline methods. We can see that our method achieves 0.970 on AUC, surpassing the performance of Multi-task (Nguyen et al., 2019), Two Branch (Masi et al., 2020), and NoiseDF (Wang and Chow, 2023) methods. Additionally, it demonstrates competitiveness with the leading performance, which is an AUC score of 0.998 attained by the MDD method (Zhao H. et al., 2021).

4.3 Cross-dataset evaluation

4.3.1 Cross-dataset evaluation

The cross-dataset generalization performance of the proposed method and comparison with the baselines are also shown in Table 1. All models are trained on the FF++ dataset and subsequently evaluated on the unseen CDF2 dataset. Results indicate that all methods suffer significant performance degradation in this challenging cross-dataset scenario when compared to the intra-dataset scenario. For instance, MDD (Zhao H. et al., 2021) declines from 0.998 to 0.674 AUC. In contrast, our DID method demonstrates impressive generalization capability, outperforming the CFFs (Yu et al., 2022) and NoiseDF (Wang and Chow, 2023) methods by margins of 4.99% (0.779 vs. 0.742) and 2.635% (0.779 vs. 0.759) respectively in terms of AUC. These results confirm the effectiveness and superiority of our framework.

4.3.2 Generalization on diffusion-generated facial forgery dataset

To further assess the generalization performance of our proposed method, we conduct deepfake detection on the diffusion-generated facial forgery dataset, which presents a significant challenge to existing detectors (Wang et al., 2023). All models are trained using the training set from the FF++ dataset and subsequently tested on the test set of DiFF (Cheng et al., 2024) dataset (which is unseen during training). DiFF comprises 23,661 pristine facial images and four kinds of high-quality forgery images [Text-to-Image (T2I), Image-to-Image (I2I), Face Swapping (FS), and Face Editing (FE), with a total of 537,466] generated by thirteen state-of-the-art diffusion methods.

Table 2 shows the performance comparisons with three widely used detectors [their results are cited from Cheng et al. (2024)]. It is clear from the table that our proposed DID method achieves remarkable generalization performance on the diffusion-generated facial forgery dataset. DID exceeds the competitor detectors by a large margin on the facial forgery dataset generated with T2I, I2I and FE diffusion techniques. For instance, DID exhibits advantages over DIRE (Wang et al., 2023) which is specifically designed for deepfake detection of diffusion technique-generated images by 81.45% (for T2I), 12.82% (for I2I), and 17.16% (for FE), respectively. Additionally, DID is highly competitive on the forgery dataset generated with FS diffusion technique. These results further demonstrate the superiority of our DID method.

4.4 Ablation study

4.4.1 Effect on different training datasets and backbones

To further evaluate the generalization capability of our proposed method under different cross-dataset situations and backbone architectures, we train our DID model on the DFFD dataset (Dang et al., 2020) using two different backbones (ResNet50

TABLE 3 Evaluation on DFFD to CDF2 with different backbones.

Models	AUC ↑	EER ↓
ResNet50 + BCE	0.620	0.411
ResNet50 + DID	**0.727**	**0.332**
EfficientNet-v2-L + BCE	0.716	0.344
EfficientNet-v2-L + DID	**0.763**	**0.302**

The highest results are highlighted in bold.

and EfficientNet-v2-L) and test its performance on the Celeb-DF test set. DFFD consists of real images and deepfakes generated by various methods, including FaceSwap, Deepfake, Face2Face, FaceAPP, StarGAN (Choi et al., 2018), PGGAN (Karras et al., 2018) (two versions), StyleGAN (Karras et al., 2019), and videos from Deep Face Lab. Experiments are conducted on a subset of DFFD following the protocols established in Dang et al. (2020), excluding inaccessible videos.

As shown in Table 3, DID consistently achieves performance improvement across all backbones fine-tuned with the BCE loss. Notably, it achieves a 17.26% gain in AUC (0.727 vs. 0.620) and a 19.22% improvement in EER (0.763 vs. 0.716) compared to the ResNet-50 backbone. With EfficientNet-V2-L as the backbone, the improvements are 6.56% (0.763 vs. 0.716) on AUC and 12.21% (0.302 vs. 0.344) on EER. These results demonstrate the applicability of our method across different datasets and various feature extraction backbones in the context of cross-dataset deepfake detection.

4.4.2 The effect of AUC loss

The effect of hyperparameter α in the AUC loss, as shown in Equation 9, is analyzed by training our model with varying values of $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ and presenting the AUC results in Figure 4. When the model is trained using only the AUC loss ($\alpha = 0.0$) as the deepfake classification loss function, it achieves the lowest AUC score of 0.380. On the other hand, when trained using only the BCE deepfake classification loss ($\alpha = 1.0$), the model obtains an AUC of 0.763. Interestingly, the model trained with an equal weighting of both the AUC loss and the BCE loss ($\alpha = 0.5$) delivers the highest AUC score of 0.779, indicating that the AUC loss contributes positively to improving the model's generalization performance.

4.4.3 The effect of A_{dom} and T modules

To investigate the significance of the domain attention module A_{dom} and the decorrelation learning module T, we conducted ablation experiments by training the proposed DID framework with each module removed individually. The detection performance of these models is presented in Table 4. From the results, we observe that when the domain attention module A_{dom} is excluded from the DID framework (denoted as "w/o A_{dom}" in Table 4), the AUC score drops by 2.05% (from 0.779 to 0.763) and the EER increases by 5.59% (from 0.286 to 0.302) compared to the complete DID model. Moreover, when the decorrelation learning module T is removed (denoted as "w/o T" in Table 4), the AUC decreases to 0.759, and the EER rises to 0.305, leading to a more significant performance drop than the model without A_{dom}. Specifically, the AUC score decreases by 2.57%, and the EER increases by 6.64%. These results emphasize the necessity of both the A_{dom} and T modules in ensuring the robustness and effectiveness of the DID framework.

4.4.4 Analysis of domain classification module

Figure 5 presents the confusion matrix of the domain feature classification. The matrix clearly demonstrates that the domain classification module achieves excellent accuracy in distinguishing various forgery methods. Notably, the average classification accuracy across all methods is 0.91, with the FaceSwap method being classified with the highest accuracy of 0.99. These results indicate that the domain-specific information is effectively separated from the deepfake-relevant features and successfully captured by the domain classification module. This outcome aligns with our decomposition objective and significantly enhances the deepfake detection process.

4.5 Visualization

4.5.1 Visualization of the saliency map

To provide a more intuitive understanding of our method's effectiveness, we visualize the Grad-CAM outputs of the deepfake attention map M_{df} and the domain (forgery technique) attention map M_{dom} in Figure 6. The figure shows that the activation regions of M_{df} and M_{dom} differ significantly. M_{dom} primarily highlights facial regions such as the nose, mouth, and eyes. On the other hand, M_{df} focuses on the information that remains consistent across different forgery techniques. These visualizations confirm the efficacy of our approach: the decorrelation learning module successfully encourages the disentangled components to contain distinct, independent information, as intended by the design of our method.

4.5.2 Visualization of deepfake features

Figures 7a, b present the T-SNE visualizations of the deepfake feature vectors extracted by the backbone network EfficientNet-v2-L and our DID framework, respectively. In both figures, red and blue dots represent real and fake face image features, respectively. The figures show that feature vectors from the backbone network are intermingled in the feature space, indicating poor separation between real and fake features. In contrast, the feature vectors generated by our DID framework are clearly separated in the feature space, demonstrating superior discrimination between real and fake faces. This highlights the effectiveness of DID's deepfake feature representation.

4.5.3 Visualization of domain features

Figure 7c further presents a visualization of the domain features acquired by our DID framework. As depicted in the figure, the domain features extracted from facial images generated by diverse forgery techniques are distinctly separable within the embedding space. The features corresponding to the same forgery approach are closely grouped together, while those from different methods

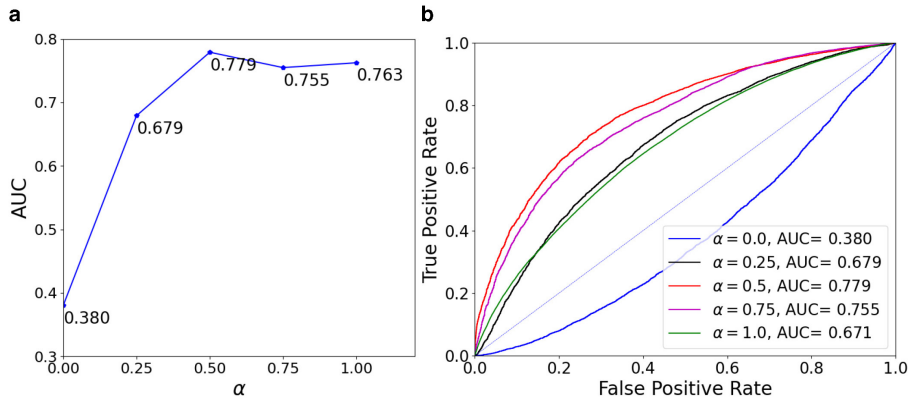

FIGURE 4
Effect of different α values (used as the balance factor between BCE and AUC loss) on the AUC score. From left to right are: **(a)** is AUC with different α values; **(b)** is ROC with different α values.

TABLE 4 Ablation study of removing ("w/o") the domain attention A_{dom} or the decorrelation learning T module from the DID framework.

Models	Modules			AUC ↑	EER ↓
	A_{df}	A_{dom}	T		
w/o A_{dom}	√	×	√	0.763	0.302
w/o T	√	√	×	0.759	0.305
DID	√	√	√	0.779	0.286

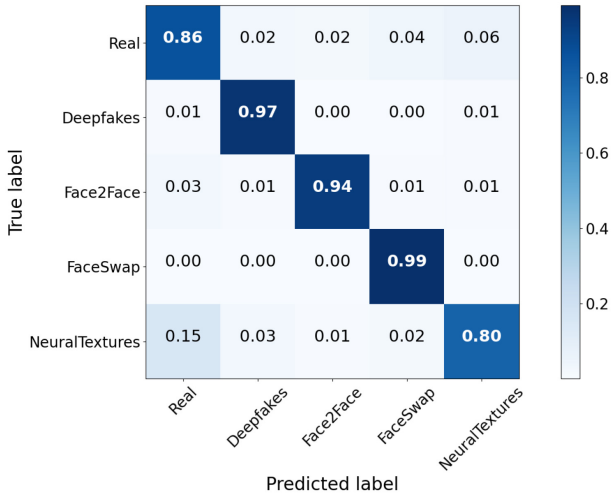

FIGURE 5
Confusion matrix visualization of domain feature classification. Each deepfake technique is recognized by the domain classification module with high accuracy (the value on the diagonal).

are spaced far apart. These observations validate that our DID framework effectively captures and separates deepfake technique-related information. It should be noted that, although I_{dom} demonstrates a remarkable ability to discriminate among different forgery techniques as illustrated in Figure 7c, this capability is confined to the forgery techniques that were presented during the training phase. In the case of encountering unfamiliar forged images, I_{dom} fails to precisely identify the forgery techniques employed in the images and is also incapable of accurately ascertaining the authenticity of the images.

5 Conclusion

In this paper, we introduce a deep information decomposition (DID) framework that decomposes facial representations in deepfake images into deepfake-related and unrelated information. The framework further refines these components to ensure clear separation, leveraging only deepfake-related features for distinguishing real from fake images. This approach enhances the deepfake detection model's robustness to irrelevant variations and improves generalization to unseen manipulation techniques. Extensive quantitative evaluations and visual analyses demonstrate the effectiveness and superiority of the proposed DID framework in cross-dataset deepfake detection.

The DID framework is designed with generalizability in mind, making it potentially applicable to other tasks such as pose-, expression-, and age-invariant face recognition, although still possesses certain limitations. Currently, all hyperparameters in the loss function require manual tuning through extensive experimental trials, which is both inefficient and suboptimal. Moreover, the domain classification module depends on access to domain-specific information from the original deepfake datasets, which is often unavailable or incomplete in real-world scenarios.

To address these limitations, several promising directions can be pursued in future work. First, the hyperparameter optimization process could be automated during training to reduce dependence on manual tuning and improve reproducibility. Second, an auxiliary module could be developed to infer or replace the need for explicit domain-specific information, thereby enhancing the framework's adaptability in real-world scenarios where such metadata is scarce or incomplete. These improvements are expected to significantly increase the practicality and robustness of the DID framework for real-time and large-scale deepfake detection.

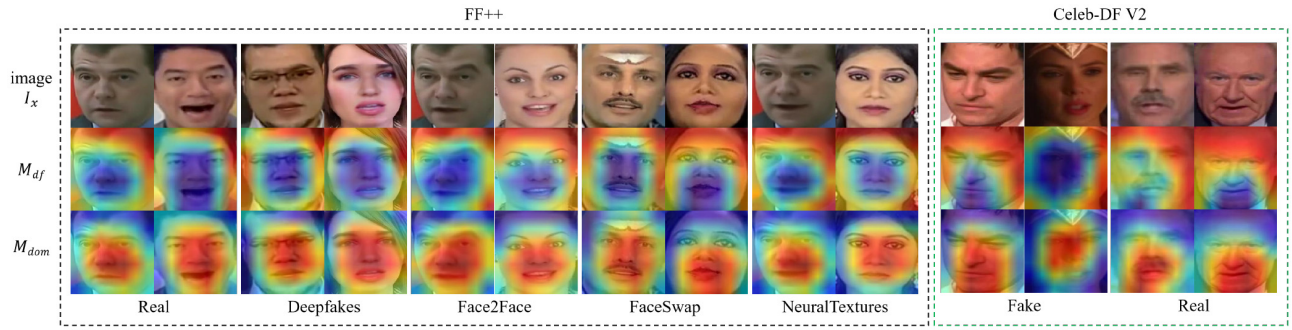

FIGURE 6
Visualization of the real/fake attention map M_{df} and the domain (forgery technique) attention map M_{dom} on the FF++ and CDF2 datasets (images adapted with permission from Rossler, A., 2019, https://www.kaggle.com/datasets/xdxd003/ff-c23, licensed under CC-BY-NC, and from Li Y., 2020, https://www.kaggle.com/datasets/reubensuju/celeb-df-v2, licensed under CC-BY). We can see that M_{dom} captures the forgery technique-related information (e.g., the forged region) while M_{df} focuses on the information invariant to forgery techniques.

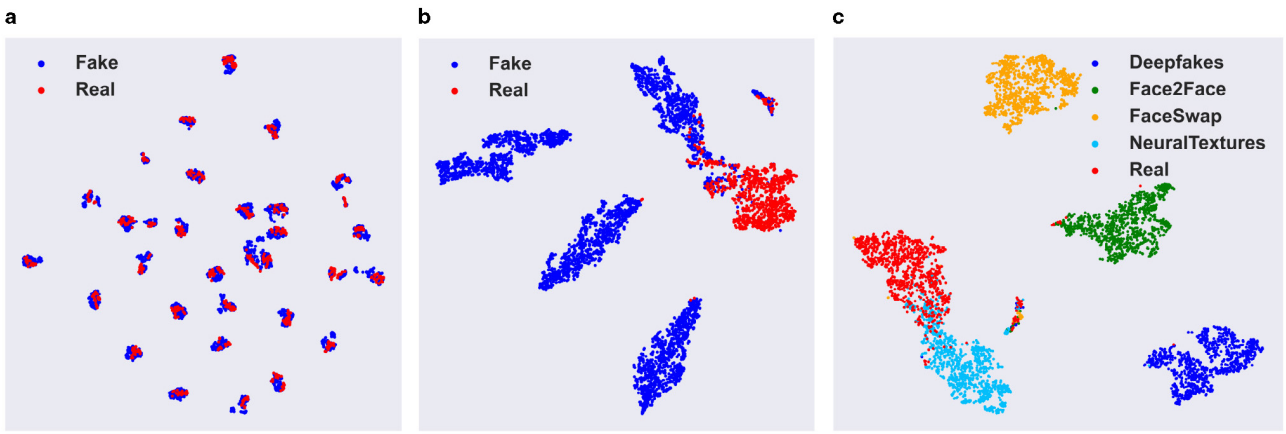

FIGURE 7
From left to right are visualizations of: **(a)** deepfake features of the EfficientNet-v2-L backbone network; **(b)** our DID framework's deepfake features; **(c)** our DID framework's domain features.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data included in this article.

Author contributions

SY: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. HG: Validation, Writing – review & editing. SH: Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. BZ: Formal analysis, Writing – review & editing. YF: Conceptualization, Data curation, Investigation, Writing – review & editing. SL: Formal analysis, Writing – review & editing. XWu: Funding acquisition, Project administration, Supervision, Writing – review & editing. XWa: Formal analysis, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Sichuan Science and Technology Program (Nos. 2024YFG0008 and 2024ZDZX0007).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. (2018). "Towards open-set identity preserving face synthesism," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Salt Lake City, UT, USA), 6713–6722. doi: 10.1109/CVPR.2018.00702

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., et al. (2018). "Mutual information neural estimation," in *Proceedings of the International Conference on Machine Learning* (Stockholm: PMLR), 531–540.

Beznosikov, A., Gorbunov, E., Berard, H., and Loizou, N. (2022). "Stochastic gradient descent-ascent: unified theory and new efficient methods," in *Proceedings of the International Conference on Artificial Intelligence and Statistics 2023, Vol. 206* (Valencia: PMLR), 172–235.

Cheng, H., Guo, Y., Wang, T., Nie, L., and Kankanhalli, M. (2024). "Diffusion facial forgery detection," in *Proceedings of the ACM International Conference on Multimedia* (New York, NY: Association for Computing Machinery), 5939–5948. doi: 10.1145/3664647.3680797

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). "Stargan: unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Salt Lake City, UT, USA), 8789–8797. doi: 10.1109/CVPR.2018.00916

Dang, H., Liu, F., Stehouwer, J., Liu, X., and Jain, A. K. (2020). "On the detection of digital face manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Seattle, WA, USA), 5781–5790. doi: 10.1109/CVPR42600.2020.00582

Dong, S., Wang, J., Liang, J., Fan, H., and Ji, R. (2022). "Explaining deepfake detection by analysing image matching," in *European Conference on Computer Vision* (Springer, Cham), 18–35. doi: 10.1007/978-3-031-19781-9_2

Ganin, Y., and Lempitsky, V. (2015). "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on Machine Learning, Vol. 37* (Lille: JMLR W&CP), 1180–1189.

Geng, Z., Cao, C., and Tulyakov, S. (2019). "3D guided fine-grained face manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Long Beach, CA, USA), 9821–9830. doi: 10.1109/CVPR.2019.01005

He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi: 10.1109/TKDE.2008.239

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., et al. (2019). "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*.

Hu, J., Wang, S., and Li, X. (2021). "Improving the generalization ability of deepfake detection via disentangled representation learning," in *IEEE International Conference on Image Processing* (IEEE: Anchorage, AK, USA), 3577–3581. doi: 10.1109/ICIP42928.2021.9506730

Huang, J., Du, C., Zhu, X., Ma, S., Nepal, S., and Xu, C. (2023). "Anti-compression contrastive facial forgery detection," in *IEEE Transactions on Multimedia, Vol. 26* (IEEE), 6166–6177. doi: 10.1109/TMM.2023.3347103

Joyce, J. M. (2011). "Kullback-leibler divergence," in *International Encyclopedia of Statistical Science* (Berlin; Heidelberg: Springer), 720–722. doi: 10.1007/978-3-642-04898-2_327

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations* (Vancouver, BC: Ithaca).

Karras, T., Laine, S., and Aila, T. (2019). "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Long Beach, CA, USA), 4401–4410. doi: 10.1109/CVPR.2019.00453

Kim, D.-K., and Kim, K.-S. (2022). "Generalized facial manipulation detection with edge region feature extraction," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision* (Waikoloa, HI: IEEE), 2828–2838. doi: 10.1109/WACV51458.2022.00284

Kinney, J. B., and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proc. Nat. Acad. Sci.* 111, 3354–3359. doi: 10.1073/pnas.1309933111

Li, J., Xie, H., Li, J., Wang, Z., and Zhang, Y. (2021). "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Nashville, TN, USA), 6458–6467. doi: 10.1109/CVPR46437.2021.00639

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., et al. (2020). "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Seattle, WA, USA), 5001–5010. doi: 10.1109/CVPR42600.2020.00505

Li, Y., Chang, M.-C., and Lyu, S. (2018). "In ICTU oculi: exposing AI created fake videos by detecting eye blinking," in *IEEE International Workshop on Information Forensics and Security* (IEEE: Hong Kong, China), 1–7. doi: 10.1109/WIFS.2018.8630787

Li, Y., and Lyu, S. (2019). "Exposing deepfake videos by detecting face warping artifacts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Long Beach, CA).

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). "Celeb-DF: a large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Seattle, WA, USA), 3207–3216. doi: 10.1109/CVPR42600.2020.00327

Liang, J., Shi, H., and Deng, W. (2022). "Exploring disentangled content information for face forgery detection," in *Proceedings of the European Conference on Computer Vision* (Springer, Cham), 128–145. doi: 10.1007/978-3-031-19781-9_8

Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., and AbdAlmageed, W. (2020). "Two-branch recurrent network for isolating deepfakes in videos," in *Proceedings of the European Conference on Computer Vision* (Springer, Cham), 667–684. doi: 10.1007/978-3-030-58571-6_39

Menéndez, M., Pardo, J., Pardo, L., and Pardo, M. (1997). The jensen-shannon divergence. *J. Franklin Inst.* 334, 307–318. doi: 10.1016/S0016-0032(96)00063-4

Nadimpalli, A. V., and Rattani, A. (2022). "On improving cross-dataset generalization of deepfake detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: New Orleans, LA, USA), 91–99. doi: 10.1109/CVPRW56347.2022.00019

Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., and Fierrez, J. (2020). Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE J. Sel. Top. Signal Process.* 14, 1038–1048. doi: 10.1109/JSTSP.2020.3007250

Nguyen, H. H., Fang, F., Yamagishi, J., and Echizen, I. (2019). "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *IEEE International Conference on Biometrics Theory, Applications and Systems* (Tampa, FL: IEEE), 1–8. doi: 10.1109/BTAS46853.2019.9185974

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., et al. (2022). Deep learning for deepfakes creation and detection: a survey. *Comput. Vis. Image Underst.* 223:103525. doi: 10.1016/j.cviu.2022.103525

Pu, W., Hu, J., Wang, X., Li, Y., Hu, S., Zhu, B., et al. (2022). Learning a deep dual-level network for robust deepfake detection. *Pattern Recognit.* 130:108832. doi: 10.1016/j.patcog.2022.108832

Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. (2020). "Thinking in frequency: face forgery detection by mining frequency-aware clues," in *Proceedings of the European Conference on Computer Vision* (Springer, Cham), 86–103. doi: 10.1007/978-3-030-58610-2_6

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). "Faceforensics++: learning to detect manipulated facial images," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE: Seoul, South Korea), 1–11. doi: 10.1109/ICCV.2019.00009

Shiohara, K., and Yamasaki, T. (2022). "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: New Orleans, LA, USA), 18720–18729. doi: 10.1109/CVPR52688.2022.01816

Tan, M., and Le, Q. (2021). "Efficientnetv2: smaller models and faster training," in *Proceedings of the International Conference on Machine Learning, Vol. 139* (PMLR), 10096–10106.

Thies, J., Zollhófer, M., and Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph*. 38, 1–12. doi: 10.1145/3306346.3323035

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). "Face2face: real-time face capture and reenactment of rgb videos," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2387–2395. doi: 10.1145/2929464.2929475

Tran, L., Yin, X., and Liu, X. (2017). "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1415–1424. doi: 10.1109/CVPR.2017.141

Wang, H., Gong, D., Li, Z., and Liu, W. (2019). "Decorrelated adversarial learning for age-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Long Beach, CA, USA), 3527–3536. doi: 10.1109/CVPR.2019.00364

Wang, T., and Chow, K. P. (2023). Noise based deepfake detection via multi-head relative-interaction. *Proc. AAAI Conf. Artif. Intell.* 37, 14548–14556. doi: 10.1609/aaai.v37i12.26701

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., et al. (2023). Dire for diffusion-generated image detection. *In Proceedings of the IEEE International Conference on Computer Vision* (IEEE: Paris, France), 22445–22455. doi: 10.1109/ICCV51070.2023.02051

Wu, X., Huang, H., Patel, V. M., He, R., and Sun, Z. (2019). Disentangled variational representation for heterogeneous face recognition. *Proc. AAAI Conf. Artif. Intell.* 33, 9005–9012. doi: 10.1609/aaai.v33i01.33019005

Yan, L., Dodier, R. H., Mozer, M., and Wolniewicz, R. H. (2003). "Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic," in *Proceedings of the International Conference on Machine Learning* (Menlo Park, CA: AAAI Press), 848–855.

Yan, Z., Zhang, Y., Fan, Y., and Wu, B. (2023). "UCF: uncovering common features for generalizable deepfake detection," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE: Paris, France), 22412–22423. doi: 10.1109/ICCV51070.2023.02048

Yang, S., Yang, X., Lin, Y., Cheng, P., Zhang, Y., and Zhang, J. (2021). "Heterogeneous face recognition with attention-guided feature disentangling," in *Proceedings of the 29th ACM International Conference on Multimedia* (New York, NY: Association for Computing Machinery), 4137–4145. doi: 10.1145/3474085.3475546

Yang, X., Li, Y., and Lyu, S. (2019). "Exposing deep fakes using inconsistent head poses," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE: Brighton, UK), 8261–8265. doi: 10.1109/ICASSP.2019.8683164

Yin, Z., Wang, J., Xiao, Y., Zhao, H., Li, T., Zhou, W., et al. (2024). Improving deepfake detection generalization by invariant risk minimization. *IEEE Trans. Multimedia*. 26, 6785–6798. doi: 10.1109/TMM.2024.3355651

Yu, M., Li, H., Yang, J., Li, X., Li, S., and Zhang, J. (2024). FDML: feature disentangling and multi-view learning for face forgery detection. *Neurocomputing* 572:127192. doi: 10.1016/j.neucom.2023.127192

Yu, P., Fei, J., Xia, Z., Zhou, Z., and Weng, J. (2022). Improving generalization by commonality learning in face forgery detection. *IEEE Trans. Inf. Forensics Security* 17, 547–558. doi: 10.1109/TIFS.2022.3146781

Yu, Y., Ni, R., Yang, S., Zhao, Y., and Kot, A. C. (2023). Narrowing domain gaps with bridging samples for generalized face forgery detection. *IEEE Trans. Multimed.* 26, 3405–3417. doi: 10.1109/TMM.2023.3310341

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* 23, 1499–1503. doi: 10.1109/LSP.2016.2603342

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. (2021). "Multi-attentional deepfake detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE: Nashville, TN, USA), 2185–2194. doi: 10.1109/CVPR46437.2021.00222

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., and Xia, W. (2021). "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE: Montreal, QC, Canada), 15023–15033. doi: 10.1109/ICCV48922.2021.01475

Zhu, C., Zhang, B., Yin, Q., Yin, C., and Lu, W. (2024). Deepfake detection via inter-frame inconsistency recomposition and enhancement. *Pattern Recognit.* 147:110077. doi: 10.1016/j.patcog.2023.110077